Appliquer le framework AWS Well-Architected pour Amazon Neptune

AWS Conseils prescriptifs

AWS Conseils prescriptifs: Appliquer le framework AWS Well-Architected pour Amazon Neptune

Table of Contents

Appliquer le framework AWS Well-Architected pour Amazon Neptune

Amazon Web Services (contributeurs)

Janvier 2026 (historique du document)

Vous pouvez créer des solutions basées sur des graphiques sur Amazon Web Services (AWS) à l'aide d'Amazon Neptune. Ce guide fournit des conseils prescriptifs pour appliquer les principes d'AWS Well-Architected Framework lorsque vous planifiez votre déploiement de Neptune.

Le AWS Well-Architected Framework vous aide à créer des infrastructures sécurisées, performantes, résilientes et efficaces pour une variété d'applications et de charges de travail. Il fournit également une approche cohérente vous permettant d'évaluer les architectures et de mettre en œuvre des conceptions évolutives.

Le AWS Well-Architected Framework repose sur les six piliers suivants :

- Excellence opérationnelle
- Sécurité
- Fiabilité
- Efficacité des performances
- Optimisation des coûts
- Durabilité

Ce guide fournit des informations sur les AWS piliers de conception et les meilleures pratiques du Well-Architected Framework, ainsi que les points à prendre en compte lors du déploiement de Neptune sur. AWS

Public visé

Ce guide est destiné aux ingénieurs de données, aux architectes de solutions et aux analystes de données qui conçoivent et mettent en œuvre des solutions utilisant des graphiques sur AWS.

Objectifs

Ce guide peut vous aider, vous et votre organisation, à effectuer les tâches suivantes :

- Choisissez parmi les options de déploiement et les langages de requête pris en charge, en fonction de votre cas d'utilisation et de vos modèles de requête.
- Suivez les AWS modèles de conception de Well-Architected qui vous aideront à améliorer la résilience et la sécurité.
- Concevez vos requêtes pour des performances optimales et des économies de coûts.
- Découvrez comment être efficace sur le plan opérationnel lors de la gestion de votre cluster Neptune en production.

Pilier d’excellence opérationnelle

Le pilier de l'excellence opérationnelle du AWS Well-Architected Framework se concentre sur le fonctionnement et le suivi des systèmes, ainsi que sur l'amélioration continue des processus et des procédures. Cela inclut la capacité de soutenir le développement et d'exécuter efficacement les charges de travail, de mieux comprendre leur fonctionnement et d'améliorer en permanence les processus et procédures de support afin de créer de la valeur commerciale. Vous pouvez réduire la complexité opérationnelle grâce à des charges de travail autoréparables, qui détectent et corrigent la plupart des problèmes sans intervention humaine. Vous pouvez atteindre cet objectif en suivant les meilleures pratiques décrites dans cette section. Utilisez les métriques et les mécanismes Amazon Neptune pour réagir correctement lorsque votre charge de travail s'écarte du comportement attendu. APIs

Cette discussion sur le pilier de l'excellence opérationnelle met l'accent sur les domaines clés suivants :

- Infrastructure en tant que code (IaC)
- Gestion des modifications
- Stratégies de résilience
- Gestion des incidents
- Rapports d'audit pour la conformité
- Journalisation et surveillance

Automatisez le déploiement en utilisant une approche IaC

Les meilleures pratiques pour automatiser le déploiement sur Neptune à l'aide d'IaC sont les suivantes :

- Appliquez l'infrastructure en tant que code (IaC) pour déployer des clusters Neptune dans la mesure du possible. Pour une configuration cohérente de l'environnement AWS Cloud Development Kit (AWS CDK), utilisez un AWS CloudFormationmodèle ou HashiCorp Terraform pour créer toutes les ressources requises pour votre cluster.
- Automatisez les procédures opérationnelles de Neptune, telles que le redimensionnement des instances, l'ajout ou la suppression de répliques de lecture ou le basculement manuel sur des tables globales, dans la mesure du possible.

- Stockez les chaînes de connexion en dehors de votre client. Utilisez des processus d'extraction, de transformation et de chargement (ETL) pour faciliter les stratégies de blue/green déploiement, la reprise après sinistre (DR) et les migrations vers de nouveaux clusters quasiment nuls. Les chaînes de connexion peuvent être stockées dans AWS Secrets ManagerAmazon DynamoDB ou dans n'importe quel emplacement où elles peuvent être modifiées dynamiquement.
- Utilisez des balises pour ajouter des métadonnées à vos ressources Neptune et suivez leur utilisation en fonction des balises. Pour plus d'informations, consultez la section Marquage des ressources Amazon Neptune.

Effectuez des modifications fréquentes, mineures et réversibles

Les recommandations suivantes se concentrent sur de petites modifications réversibles afin de minimiser la complexité et de réduire le risque d'interruption de la charge de travail :

- Stockez les modèles et les scripts IaC dans un service de contrôle de source, tel que GitHub ou GitLab.

> **Important**
>
> Ne stockez pas les AWS informations d'identification dans le contrôle de source.

- Exigez que les déploiements iAc utilisent un service d'intégration et de livraison continues (CI/CD), tel que ou. AWS CodePipelineAWS CodeBuild Ces services compilent, testent et déploient du code dans un environnement hors production contenant un cluster Neptune éphémère avant d'avoir un impact sur votre cluster Amazon Neptune de production.
- Testez les requêtes d'infrastructure et d'application dans un environnement inférieur avant de les déployer en production. Cela minimisera le risque d'interruption et contribuera à garantir qu'ils fonctionnent bien avec votre charge de travail et votre échelle.

Anticipez les défaillances

Une infrastructure autoréparante illustre l'excellence opérationnelle en anticipant les défaillances et en tentant de résoudre les problèmes sans intervention. Les recommandations suivantes vous aideront à atteindre cette maturité avec Neptune :

- Créez un plan de surveillance qui utilise CloudWatch les métriques Amazon pour surveiller l'utilisation du processeur et de la mémoire de votre instance de base de données et comprendre les modèles d'utilisation. Créez des CloudWatch tableaux de bord et des alarmes pour les indicateurs clés et les réponses des clients Neptune figurant dans les journaux de vos applications. Pour plus d'informations sur les indicateurs d'utilisation élevée ou faible du processeur, consultez la section Utilisation CloudWatch pour surveiller les performances des instances de base de données dans Neptune dans la documentation de Neptune.

  Si vous rencontrez fréquemment out-of-memory des exceptions dans vos requêtes, pensez à réduire le nombre total de nœuds traversés par votre requête ou essayez d'utiliser une instance de la `X2` famille, dont le RAM-to-CPU ratio est plus élevé.
- Définissez des notifications pour surveiller l'état du cluster Neptune. Par exemple, `BufferCacheHitRatio` il doit être constamment élevé (supérieur à 99,9 %), alors qu'`MainRequestQueuePendingRequests`il doit être constamment faible (idéalement 0, mais cela dépend de vos besoins et de votre tolérance de latence).
- Envisagez d'utiliser des répliques de lecture pour atteindre une haute disponibilité au sein de Neptune. Vous devez disposer d'au moins deux répliques de lecture situées dans des zones de disponibilité différentes de celles de l'instance du rédacteur afin de garantir qu'une instance est toujours disponible pour répondre aux requêtes de lecture lors d'un événement de basculement.
- Adaptez automatiquement les répliques de lecture en fonction des indicateurs d'utilisation. Pour plus d'informations, consultez Dimensionnement automatique du nombre de répliques dans un cluster de base de données Amazon Neptune.
- Testez le basculement pour votre instance de base de données afin de connaître la durée du processus pour votre cas d'utilisation.
- Si votre application doit survivre à une Région AWS panne complète, envisagez d'utiliser des bases de données mondiales dans le cadre de vos plans de reprise après sinistre.

Tirez les leçons de toutes les défaillances opérationnelles

Une infrastructure d'autoréparation est un effort à long terme qui se développe par itérations lorsque de rares problèmes surviennent ou que les réponses ne sont pas aussi efficaces que prévu. L'adoption des pratiques suivantes permet de se concentrer sur cet objectif :

- Favorisez l'amélioration en tirant les leçons de tous les échecs.

- Partagez ce que vous avez appris au sein des équipes et de l'organisation. Si plusieurs équipes d'une organisation utilisent Neptune, créez un salon de discussion ou un groupe d'utilisateurs commun pour partager les connaissances et les meilleures pratiques.

Utilisez les fonctionnalités de journalisation pour surveiller les activités non autorisées ou anormales

Pour observer des modèles de performance et d'activité anormaux, stockez les journaux dans Amazon CloudWatch Logs. Tenez compte des bonnes pratiques suivantes :

- Activez la journalisation lente des requêtes. Consultez régulièrement le journal et déterminez pourquoi certaines requêtes sont lentes. Utilisez les points de terminaison d'explication et de profilage Neptune pour Gremlin, SPARQL ou OpenCypher pour mieux comprendre pourquoi ces requêtes sont lentes.
- Activez les journaux d'audit Neptune et consultez-les régulièrement pour détecter tout accès non autorisé ou toute anomalie.
- Si vous utilisez la journalisation lente des requêtes ou la journalisation des audits, activez la publication dans Logs. CloudWatch Cela vous évitera de manquer d'espace disque sur les instances. Les instances Neptune ont une capacité de stockage des journaux limitée et remplacent les anciens fichiers journaux lorsque l'espace des journaux est dépassé. CloudWatch Les journaux permettent de conserver les journaux à long terme. Les fonctionnalités de surveillance améliorées de CloudWatch Logs amélioreront votre capacité à interroger les journaux et à diagnostiquer les problèmes.
- Pour améliorer les outils d'analyse de vos journaux d'audit, vous pouvez configurer un cluster de base de données Neptune pour publier les données des journaux d'audit dans un groupe de journaux dans CloudWatch Logs. Avec CloudWatch Logs, vous pouvez effectuer une analyse en temps réel des données du journal, l'utiliser CloudWatch pour créer des alarmes et afficher les métriques, et utiliser CloudWatch les journaux pour stocker vos enregistrements de journal dans un stockage hautement durable. Pour plus d'informations, consultez Publier les journaux Neptune sur Amazon CloudWatch Logs.
- Neptune prend en charge la journalisation des actions du plan de contrôle à l'aide de. AWS CloudTrail Pour plus d'informations, consultez la section Journalisation des appels d'API Amazon Neptune avec. AWS CloudTrail

Pilier de sécurité

La sécurité du cloud AWS est la priorité absolue. En tant que AWS client, vous bénéficiez d'un centre de données et d'une architecture réseau conçus pour répondre aux exigences des entreprises les plus sensibles en matière de sécurité.

La sécurité est une responsabilité partagée entre vous AWS et vous. Le modèle de responsabilité partagée décrit ceci en tant que sécurité du cloud et sécurité dans le cloud :

- Sécurité du cloud : AWS est chargée de protéger l'infrastructure qui s'exécute Services AWS dans le AWS Cloud. AWS vous fournit également des services que vous pouvez utiliser en toute sécurité. Des auditeurs tiers testent et vérifient régulièrement l'efficacité de AWS la sécurité dans le cadre des programmes de AWS conformité. Pour en savoir plus sur les programmes de conformité qui s'appliquent à Amazon Neptune, consultez Services AWS concernés par le programme de conformité.
- Sécurité dans le cloud — Votre responsabilité est déterminée par Service AWS ce que vous utilisez. Vous êtes également responsable d'autres facteurs, y compris la sensibilité de vos données, les exigences de votre entreprise, et la législation et la réglementation applicables. Pour plus d'informations sur la confidentialité des données, consultez Questions fréquentes (FAQ) relatives à la confidentialité des données. Pour plus d'informations sur la protection des données en Europe, consultez le modèle de responsabilitéAWS partagée et le billet de blog sur le RGPD.

Le pilier de sécurité vous aide à comprendre comment appliquer le modèle de responsabilité partagée lors de l'utilisation de Neptune. Les rubriques suivantes expliquent comment configurer Neptune pour répondre à vos objectifs de sécurité et de conformité. Vous apprendrez également à utiliser d'autres outils Services AWS qui vous aideront à surveiller et à sécuriser vos ressources Neptune.

Le pilier de sécurité comprend les principaux domaines d'intérêt suivants :

- Sécurité des données
- Sécurité du réseau
- Authentification et autorisation

Mettre en œuvre la sécurité des données

Les fuites de données et les violations mettent vos clients en danger et peuvent avoir un impact négatif important sur votre entreprise. Les meilleures pratiques suivantes aident à protéger les données de vos clients contre toute exposition accidentelle ou malveillante :

- Les noms de clusters, les balises, les groupes de paramètres, les rôles Gestion des identités et des accès AWS (IAM) et les autres métadonnées ne doivent pas contenir d'informations confidentielles ou sensibles, car ces données peuvent apparaître dans les journaux de facturation ou de diagnostic.
- URIs ou les liens vers des serveurs externes stockés sous forme de données dans Neptune ne doivent pas contenir d'informations d'identification permettant de valider les demandes.
- Les instances chiffrées Neptune fournissent une couche supplémentaire de protection des données en contribuant à sécuriser vos données contre tout accès non autorisé au stockage sous-jacent. Vous pouvez utiliser le chiffrement Neptune pour renforcer la protection des données de vos applications déployées dans le cloud. Vous pouvez également utiliser le chiffrement Neptune pour répondre aux exigences de conformité relatives aux données au repos.

  Pour activer le chiffrement pour une nouvelle instance de base de données Neptune, choisissez Oui dans la section Activer le chiffrement de la console Neptune (sélectionnée par défaut) ou en définissant la propriété dans. AWS::Neptune::DBCluster::StorageEncrypted CloudFormationSi le chiffrement est activé, Neptune utilisera par défaut la clé gérée AWS Amazon Relational Database Service (Amazon RDS), ou vous pouvez créer une clé gérée par le client. Pour plus d'informations sur la création d'une instance de base de données Neptune, consultez la section Création d'un nouveau cluster de base de données Neptune. Pour plus de détails, consultez la section Chiffrement des ressources Neptune au repos. Vos instantanés automatisés et manuels utilisent le même chiffrement que celui que vous avez sélectionné pour votre cluster Neptune.
- Lorsque vous utilisez les langages SPARQL et OpenCypher, appliquez des techniques de validation et de paramétrage des entrées appropriées pour empêcher l'injection de code SQL et d'autres formes d'attaques. Évitez de créer des requêtes qui utilisent la concaténation de chaînes avec des entrées fournies par l'utilisateur. Utilisez des requêtes paramétrées ou des instructions préparées pour transmettre en toute sécurité les paramètres d'entrée à la base de données de graphes. Pour plus d'informations, consultez Exemples de requêtes paramétrées OpenCypher et SPARQL Injection Defense.

- Pour le langage Gremlin, utilisez des variantes de langage Gremlin au lieu de transmettre directement des scripts Gremlin basés sur des chaînes afin d'éviter d'éventuels problèmes d'injection.

Sécurisez vos réseaux

Un cluster de base de données Amazon Neptune ne peut être créé que dans un cloud privé virtuel (VPC) sur. AWS Jusqu'à la version 1.4.6.0 de Neptune, les points de terminaison du cluster de base de données Neptune n'étaient accessibles qu'au sein de ce VPC. À partir de Neptune 1.4.6.0 et versions ultérieures, les instances de Neptune peuvent être configurées pour être accessibles au public sur Internet. Il est recommandé d'utiliser cette fonctionnalité uniquement dans des environnements hors production afin de simplifier l'accès à Neptune pour vos développeurs (bien que l'authentification IAM soit toujours requise pour permettre l'accessibilité publique). Si l'accessibilité publique est activée, envisagez de définir des règles de groupe de sécurité entrant pour le port de votre base de données uniquement pour le trafic d'adresses IP connu. Dans les environnements de production ou dans les clusters contenant des données sensibles, sécurisez vos données Neptune en interdisant l'accès du public et en limitant l'accès au VPC où se trouve votre cluster de base de données Neptune. Pour plus d'informations, consultez Connexion à votre graphe Amazon Neptune.

Pour protéger vos données en transit, Neptune applique des connexions SSL via HTTPS à n'importe quel point de terminaison d'instance ou de cluster à l'aide de protocoles et de chiffrements sécurisés. Neptune fournit des certificats SSL pour vos instances de base de données Neptune. Les certificats SSL Neptune ne prennent en charge que les noms d'hôte des points de terminaison de cluster, de point de terminaison de lecteur et de point de terminaison d'instance.

Si vous utilisez un équilibreur de charge ou un serveur proxy (tel que HAProxy), vous devez utiliser la terminaison SSL et disposer de votre propre certificat SSL sur le serveur proxy. La transmission SSL ne fonctionne pas, car les certificats SSL fournis ne correspondent au nom d'hôte du serveur proxy. Pour plus d'informations sur la connexion aux points de terminaison Neptune via SSL, voir Utilisation du point de terminaison HTTP REST pour se connecter à une instance de base de données Neptune.

Mettre en œuvre l'authentification et l'autorisation

Pour contrôler qui peut effectuer des actions de gestion Neptune sur les clusters et les instances de base de données Neptune, activez l'authentification de base de données IAM et utilisez les informations d'identification IAM. Lorsque vous vous connectez à AWS l'aide d'informations d'identification IAM, votre rôle IAM doit disposer de politiques IAM qui accordent les autorisations

requises pour effectuer les opérations de gestion de Neptune. Assurez-vous de suivre le principe du moindre privilège, en n'accordant que les autorisations requises pour effectuer une tâche. Pour plus d'informations, consultez les sections Utilisation de différents types de politiques IAM pour contrôler l'accès à Neptune et Authentification IAM à l'aide d'informations d'identification temporaires.

Pour contrôler qui peut se connecter à un cluster Neptune et interroger les données, vous pouvez utiliser IAM pour vous authentifier auprès de votre instance ou de votre cluster de base de données Neptune. Si vous activez l'authentification IAM dans un cluster de base de données Neptune, toute personne accédant au cluster de base de données doit d'abord être authentifiée. Pour plus d'informations, voir Activation de l'authentification de base de données IAM dans Neptune pour connaître les étapes à suivre pour activer l'authentification IAM.

Lorsque l'authentification de base de données IAM est activée, chaque demande doit être signée à l'aide d' AWS Signature Version 4. Pour comprendre comment envoyer des demandes signées à tous les points de terminaison Neptune lorsque l'authentification IAM est activée, voir Connexion et signature avec AWS signature version 4. De nombreuses bibliothèques et outils, tels que awscurl, prennent déjà en charge la version 4 de AWS Signature.

Pour interagir avec d'autres utilisateurs Services AWS, Amazon Neptune utilise des rôles liés au service IAM. Un rôle lié à un service est un type unique de rôle IAM lié directement à Neptune. Les rôles liés au service sont prédéfinis par Neptune et incluent toutes les autorisations dont le service a besoin pour appeler d'autres Services AWS personnes en votre nom. Pour plus d'informations, consultez la section Utilisation de rôles liés à un service pour Neptune.

Pilier de fiabilité

Le pilier de fiabilité englobe la capacité d'une charge de travail à exécuter correctement et de manière cohérente la fonction prévue lorsqu'elle est censée le faire. Cela inclut la possibilité d'exploiter et de tester la charge de travail tout au long de son cycle de vie.

Pour garantir la fiabilité d'une charge de travail, il faut commencer par choisir le bon logiciel et la bonne infrastructure. Vos choix d'architecture auront un impact sur le comportement de votre charge de travail dans tous les piliers de AWS Well-Architected. Pour des raisons de fiabilité, vous devez suivre des modèles spécifiques.

Le pilier de fiabilité se concentre sur les domaines clés suivants :

- Architecture de charge de travail, y compris les quotas de service et les modèles de déploiement
- Gestion des modifications
- Gestion des défaillances

Comprendre les quotas de service Neptune

Un volume de cluster Neptune peut atteindre une taille maximale de 128 tebioctets (TiB) dans tous les pays pris en charge, Régions AWS sauf en Chine, où le quota est de GovCloud 64 TiB.

Le quota de 128 TiB est suffisant pour stocker environ 200 à 400 milliards d'objets dans le graphique. Dans un graphe de propriétés étiqueté (LPG), un objet est un nœud, une arête ou une propriété sur un nœud ou une arête. Dans un graphe RDF (Resource Description Framework), un objet est un quad.

Pour tout cluster Neptune Serverless, vous définissez le nombre minimum et maximum d'unités de capacité Neptune (). NCUs Chaque NCU comprend 2 Gibioctets (GiB) de mémoire, le vCPU et le réseau associés. Les valeurs NCU minimale et maximale s'appliquent à toutes les instances sans serveur du cluster. La valeur NCU maximale la plus élevée que vous pouvez définir est de 128,0 NCUs et la valeur minimale la plus basse est de 1,0. NCUs Optimisez la plage NCU qui convient le mieux à votre application en observant les CloudWatch métriques `ServerlessDatabaseCapacity` d'Amazon, en capturant la plage dans laquelle vous vous trouvez le plus souvent et en corrélant les comportements ou les coûts indésirables dans cette fourchette. `NCUUtilization` Dans de nombreuses charges de travail, 1,0 NCU est un point de départ trop bas et entraîne un comportement peu fiable après des périodes d'inactivité. Si vous constatez que votre charge de travail n'évolue pas

assez rapidement, augmentez le minimum NCUs afin de fournir suffisamment de traitement pour l'augmentation initiale pendant qu'elle évolue.

Chacune Compte AWS dispose de quotas pour chaque région quant au nombre de ressources de base de données que vous pouvez créer. Ces ressources incluent des instances et des clusters de bases de données. Une fois qu'une limite a été atteinte pour une ressource, les appels supplémentaires pour créer cette ressource échouent avec une exception. Certains quotas sont des quotas souples qui peuvent être augmentés sur demande. Pour obtenir la liste des quotas partagés entre Amazon Neptune et Amazon RDS, Amazon Aurora et Amazon DocumentDB (avec compatibilité avec MongoDB), ainsi que des liens permettant de demander des augmentations de quotas lorsqu'elles sont disponibles, consultez la section Quotas dans Amazon RDS.

Comprendre les modèles de déploiement de Neptune

Dans les clusters de base de données Neptune, il existe une instance de base de données principale et jusqu'à 15 répliques de Neptune. L'instance de base de données principale prend en charge les opérations de lecture et d'écriture et effectue toutes les modifications de données sur le volume du cluster. Les répliques Neptune se connectent au même volume de stockage que l'instance de base de données principale et ne prennent en charge que les opérations de lecture. Les répliques Neptune peuvent décharger les charges de travail de lecture de l'instance de base de données principale.

Pour atteindre une haute disponibilité, utilisez des répliques de lecture. Le fait qu'une ou plusieurs instances de réplication en lecture soient disponibles dans différentes zones de disponibilité peut augmenter la disponibilité, car les répliques en lecture servent de cibles de basculement pour l'instance principale. Si l'instance Writer échoue, Neptune fait en sorte qu'une instance de réplication en lecture devienne l'instance principale. Dans ce cas, il y a une brève interruption (généralement inférieure à 30 secondes) pendant le redémarrage de l'instance promue, au cours de laquelle les demandes de lecture et d'écriture adressées à l'instance principale échouent, sauf exception. Pour une fiabilité maximale, envisagez deux répliques en lecture dans différentes zones de disponibilité. Si l'instance principale de la zone de disponibilité 1 est mise hors ligne, l'instance de la zone de disponibilité 2 est promue en instance principale, mais elle ne peut pas traiter les requêtes pendant ce temps. Une instance de la zone de disponibilité 3 est donc requise pour gérer les requêtes de lecture pendant la transition.

Si vous utilisez Neptune Serverless, les instances de lecture et d'écriture de toutes les zones de disponibilité augmenteront ou diminueront, indépendamment les unes des autres, en fonction de la charge de leur base de données. Vous pouvez définir le niveau de promotion d'une instance

de lecteur sur 0 ou 1 afin qu'il augmente ou diminue en fonction de la capacité de l'instance de rédacteur. Cela le rend prêt à prendre en charge la charge de travail actuelle à tout moment.

Si votre application est présente dans le monde entier ou nécessite un basculement multirégional, pensez à utiliser une base de données mondiale Neptune. Une base de données mondiale Amazon Neptune couvre plusieurs bases de données Régions AWS, ce qui permet des lectures globales à faible latence et une restauration rapide dans les rares cas où une panne affecte l'ensemble d'une base de données. Région AWS Une base de données globale Neptune se compose d'un cluster de bases de données principal dans une région et d'un maximum de cinq clusters de bases de données secondaires dans différentes régions.

Gérez et dimensionnez les clusters Neptune

Vous pouvez utiliser l'auto-scaling Neptune pour ajuster automatiquement le nombre de répliques Neptune dans un cluster de base de données afin de répondre à vos exigences en matière de connectivité et de charge de travail en fonction des seuils d'utilisation du processeur. Grâce à l'auto-scaling, votre cluster de base de données Neptune peut gérer des augmentations soudaines de la charge de travail. Lorsque la charge de travail diminue, l'auto-scaling supprime les répliques inutiles afin que vous ne payiez pas pour la capacité inutilisée. Sachez que le démarrage d'une nouvelle instance peut prendre jusqu'à 15 minutes. L'auto-scaling ne suffit donc pas à lui seul à faire face à l'évolution rapide de la demande.

Vous ne pouvez utiliser l'auto-scaling qu'avec un cluster de base de données Neptune qui possède déjà une instance d'écriture principale et au moins une instance de reproduction en lecture (voir Clusters et instances de base de données Amazon Neptune). En outre, toutes les instances de réplica en lecture du cluster doivent être dans un état disponible. Si une réplique en lecture est dans un état autre que disponible, l'auto-scaling de Neptune ne fait rien tant que toutes les répliques en lecture du cluster ne sont pas disponibles.

Si la demande évolue rapidement, pensez à utiliser des instances sans serveur. Les instances sans serveur peuvent évoluer verticalement sur de courtes périodes tandis que l'auto-scaling s'adapte horizontalement sur de longues périodes. Cette configuration offre une évolutivité optimale car les instances sans serveur évoluent verticalement tandis que l'auto-scaling instancie de nouvelles répliques de lecture afin de gérer la charge de travail au-delà de la capacité maximale d'une seule instance sans serveur. Pour plus d'informations sur le dimensionnement de la capacité d'Amazon Neptune Serverless, consultez la section Dimensionnement de capacité dans un cluster de base de données Neptune Serverless.

Si vos besoins en matière de dimensionnement évoluent à des moments prévisibles, vous pouvez planifier des modifications des instances minimales, maximales et seuils afin de mieux répondre à ces besoins changeants. N'oubliez pas de planifier les événements à grande échelle au moins 15 minutes à l'avance pour permettre à ces instances d'être mises en ligne en cas de besoin.

Vous gérez votre configuration de base de données dans Amazon Neptune à l'aide de paramètres dans un groupe de paramètres. Les groupes de paramètres servent de conteneurs pour les valeurs de configuration de moteur qui sont appliquées à une ou plusieurs instances de base de données. Lorsque vous modifiez des paramètres de cluster dans des groupes de paramètres, comprenez la différence entre les paramètres statiques et dynamiques, et apprenez comment et quand ils sont appliqués. Utilisez le point de terminaison d'état pour voir la configuration actuellement appliquée.

Gestion des sauvegardes et des événements de basculement

Neptune sauvegarde automatiquement le volume de votre cluster et conserve les données sauvegardées pendant toute la durée de la période de conservation des sauvegardes. Les sauvegardes Neptune étant continues et incrémentielles, vous pouvez rapidement opérer une restauration à un point quelconque de la période de rétention des sauvegardes. Vous pouvez spécifier une période de rétention des sauvegardes de 1 à 35 jours lorsque vous créez ou modifiez un cluster de bases de données.

Pour conserver une sauvegarde au-delà de la période de conservation des sauvegardes, vous pouvez également prendre un instantané des données de votre volume de cluster. Le stockage des instantanés est soumis aux frais standard de stockage pour Neptune.

Lorsque vous créez un instantané Amazon Neptune d'un cluster de bases de données, Neptune crée un instantané du volume de stockage du cluster, en sauvegardant toutes ses données, et pas seulement des instances individuelles. Vous pourrez ultérieurement créer un cluster de bases de données en effectuant une restauration à partir de cet instantané. Lorsque vous restaurez le cluster de base de données, vous fournissez le nom du snapshot du cluster de base de données à partir duquel effectuer la restauration, puis vous fournissez un nom pour le nouveau cluster de base de données créé par la restauration.

Testez la façon dont votre système réagit aux événements de basculement. Utilisez l'API Neptune pour forcer un événement de basculement. Le redémarrage avec basculement est utile lorsque vous souhaitez simuler la défaillance d'une instance de base de données à des fins de test ou pour des opérations de restauration dans la zone de disponibilité d'origine après un basculement. Pour plus d'informations, consultez Configuration et gestion d'un déploiement multi-AZ. Lorsque vous

redémarrez une instance de rédacteur de base de données, elle bascule vers la réplique de secours. Le redémarrage d'un réplica Neptune ne déclenche pas de basculement.

Concevez vos clients dans une optique de fiabilité. Testez leur comportement lors d'événements de basculement. Implémentez une logique de nouvelle tentative dans votre client avec une logique d'interruption exponentielle. Des exemples de code implémentant cette logique sont disponibles dans la documentation sous les exemples de AWS Lambda fonctions pour Amazon Neptune.

Envisagez de l'utiliser AWS Backupsi vous avez des exigences communes en matière de sauvegarde que vous appliquez à plusieurs moteurs de base de données.

Pilier d’efficacité des performances

Le pilier de l'efficacité des performances du AWS Well-Architected Framework se concentre sur la manière d'optimiser les performances lors de l'ingestion ou de l'interrogation de données. L'optimisation des performances est un processus progressif et continu comprenant les éléments suivants :

- Confirmation des exigences commerciales
- Mesurer les performances de la charge de travail
- Identification des composants sous-performants
- Optimisation des composants pour répondre aux besoins de votre entreprise

Le pilier relatif à l'efficacité des performances fournit des directives spécifiques aux cas d'utilisation qui peuvent aider à identifier le modèle de données graphique et les langages de requête appropriés à utiliser. Il inclut également les meilleures pratiques à suivre lors de l'ingestion et de la consommation de données dans Amazon Neptune.

Le pilier de l'efficacité en matière de performance se concentre sur les domaines clés suivants :

- Modélisation de graphes
- Optimisation des requêtes
- Dimensionnement correct du cluster
- Optimisation de l'écriture

Comprendre la modélisation de graphes

Comprenez la différence entre les modèles Labeled Property Graph (LPG) et Resource Description Framework (RDF). Dans la plupart des cas, c'est une question de préférence. Il existe toutefois plusieurs cas d'utilisation où un modèle est mieux adapté que l'autre. Si vous avez besoin de connaître le chemin reliant deux nœuds de votre graphe, choisissez LPG. Si vous souhaitez fédérer les données entre des clusters Neptune ou d'autres magasins triples de graphes, choisissez RDF.

Si vous créez une application logicielle en tant que service (SaaS) ou une application nécessitant une mutualisation, envisagez d'intégrer la séparation logique des locataires dans votre modèle de données au lieu d'avoir un locataire pour chaque cluster. Pour obtenir ce type de conception,

vous pouvez utiliser des graphes nommés et des stratégies d'étiquetage SPARQL, telles que l'ajout d'identifiants clients aux étiquettes ou l'ajout de paires clé-valeur de propriété représentant les identifiants des locataires. Assurez-vous que votre couche client injecte ces valeurs pour conserver cette séparation logique. Pour plus d'informations sur les recommandations relatives à la mutualisation, consultez les instructions relatives à la mutualisation des bases de données Amazon ISVs Neptune.

Les performances de vos requêtes dépendent du nombre d'objets graphiques (nœuds, arêtes, propriétés) qui doivent être évalués lors du traitement de votre requête. Le modèle graphique peut donc avoir un impact significatif sur les performances de votre application. Utilisez des étiquettes granulaires lorsque cela est possible et ne stockez que les propriétés dont vous avez besoin pour déterminer le chemin ou filtrer. Pour améliorer les performances, pensez à précalculer certaines parties de votre graphe, par exemple en créant des nœuds de synthèse ou des arêtes plus directes reliant des chemins communs.

Essayez d'éviter de naviguer entre des nœuds présentant un nombre anormalement élevé d'arêtes portant la même étiquette. Ces nœuds ont souvent des milliers d'arêtes (alors que le nombre d'arêtes de la plupart des nœuds se chiffre en dizaines). Il en résulte une complexité de calcul et de données beaucoup plus élevée. Ces nœuds peuvent ne pas poser de problème dans certains modèles de requêtes, mais nous vous recommandons de modéliser vos données différemment pour éviter cela, en particulier si vous devez naviguer entre les nœuds comme étape intermédiaire. Vous pouvez utiliser les journaux de requêtes lentes pour identifier les requêtes qui naviguent entre ces nœuds. Vous observerez probablement des indicateurs de latence et d'accès aux données bien supérieurs à vos modèles de requête habituels, en particulier si vous utilisez le mode débogage.

Utilisez un nœud déterministe IDs pour les nœuds et les arêtes si votre cas d'utilisation le permet au lieu d'utiliser Neptune pour attribuer des valeurs GUID aléatoires pour. IDs L'accès aux nœuds par identifiant est la méthode la plus efficace.

Optimisation des requêtes

Les langages OpenCypher et Gremlin peuvent être utilisés de manière interchangeable sur les modèles GPL. Si les performances sont une préoccupation majeure, envisagez d'utiliser les deux langages de manière interchangeable, car l'un peut être plus performant que l'autre pour des modèles de requête spécifiques.

Neptune est en train de passer à son moteur de requête alternatif (DFE). OpenCypher ne fonctionne que sur le DFE, mais les requêtes Gremlin et SPARQL peuvent éventuellement être configurées

pour s'exécuter sur le DFE à l'aide d'annotations de requête. Pensez à tester vos requêtes avec le DFE activé et à comparer les performances de votre modèle de requête lorsque vous n'utilisez pas le DFE.

Neptune est optimisé pour les requêtes de type transactionnel qui démarrent sur un seul nœud ou un ensemble de nœuds et se déploient à partir de là, plutôt que pour les requêtes analytiques qui évaluent le graphe dans son intégralité. Pour vos charges de travail de requêtes analytiques, utilisez Neptune Analytics. Neptune Analytics est le choix idéal pour les charges de travail d'investigation, d'exploration ou de science des données qui nécessitent une itération rapide pour le traitement des données, des analyses et des algorithmes. Il peut également effectuer une recherche vectorielle sur des données graphiques et charger des données directement depuis votre instance de base de données Neptune. Si Neptune Analytics ne répond pas à vos besoins, vous pouvez également envisager de créer un AWS SDK pour Pandas ou d'utiliser neptune-export en combinaison avec Amazon EMR. AWS Glue

Pour identifier les inefficiences et les goulots d'étranglement dans vos modèles et requêtes, utilisez le `profile` et `explain` APIs pour chaque langage de requête afin d'obtenir des explications détaillées sur le plan des requêtes et les métriques des requêtes. Pour plus d'informations, consultez le profil Gremlin, OpenCypher explain et SPARQL explain.

Comprenez vos modèles de requêtes. Si le nombre d'arêtes distinctes dans un graphique devient important, la stratégie d'accès par défaut de Neptune peut devenir inefficace. Les requêtes suivantes peuvent s'avérer peu efficaces :

- Requêtes qui naviguent vers l'arrière sur les arêtes lorsqu'aucune étiquette de bord n'est spécifiée.
- Des clauses qui utilisent ce même modèle en interne, comme `.both()` dans Gremlin, ou des clauses qui suppriment des nœuds dans n'importe quelle langue (ce qui nécessite de supprimer les arêtes entrantes sans connaître les étiquettes).
- Requêtes qui accèdent aux valeurs des propriétés sans spécifier d'étiquettes de propriétés. Ces requêtes peuvent devenir très inefficaces. Si cela correspond à votre modèle d'utilisation, pensez à activer l'index OSGP (objet, sujet, graphe, prédicat).

Utilisez la journalisation lente des requêtes pour identifier les requêtes lentes. La lenteur des requêtes peut être due à des plans de requêtes non optimisés ou à un nombre inutilement élevé de recherches d'index, ce qui peut augmenter les coûts. I/O Les points de terminaison d'explication et de profilage Neptune pour Gremlin, SPARQL ou OpenCypher peuvent vous aider à comprendre pourquoi ces requêtes sont lentes. Les causes peuvent inclure les suivantes :

- Les nœuds présentant un nombre anormalement élevé d'arêtes par rapport à la moyenne des nœuds du graphe (par exemple, des milliers au lieu de dizaines) peuvent accroître la complexité du calcul et, par conséquent, prolonger la latence et augmenter la consommation de ressources. Déterminez si ces nœuds sont correctement modélisés ou si les modèles d'accès peuvent être améliorés afin de réduire le nombre de tronçons à franchir.
- Les requêtes non optimisées contiendront un avertissement indiquant que des étapes spécifiques ne sont pas optimisées. La réécriture de ces requêtes pour utiliser des étapes optimisées peut améliorer les performances.
- Les filtres redondants peuvent entraîner des recherches d'index inutiles. De même, les modèles redondants peuvent entraîner des recherches d'index dupliquées qui peuvent être optimisées en améliorant la requête (voir `Index Operations - Duplication ratio` dans la sortie du profil).
- Certaines langues, telles que le gremlin, n'utilisent pas de valeurs numériques fortement saisies et utilisent plutôt la promotion de type. Par exemple, si la valeur est 55, Neptune recherche des valeurs entières, longues, flottantes et autres types numériques équivalents à 55. Cela se traduit par des opérations supplémentaires. Si vous savez à l'avance que vos types correspondent, vous pouvez éviter cela en utilisant un indice de requête.
- Votre modèle de graphe peut avoir un impact important sur les performances. Envisagez de réduire le nombre d'objets à évaluer en utilisant des étiquettes plus détaillées ou en précalculant les raccourcis vers des trajectoires linéaires à sauts multiples.

Si l'optimisation des requêtes à elle seule ne vous permet pas d'atteindre vos exigences en matière de performances, envisagez d'utiliser diverses techniques de mise en cache avec Neptune pour répondre à ces exigences.

Les performances de Neptune ne cessent de s'améliorer à chaque version. Consultez les notes de publication pour connaître les détails des améliorations apportées à chaque version. Envisagez de planifier des mises à jour régulières de votre cluster de base de données Neptune afin d'obtenir des performances optimales. Les nouvelles versions prennent également en charge les nouvelles instances. Envisagez de passer à la version 1.4.5.0 ou à une version ultérieure pour pouvoir utiliser les `r8g` instances. Pour plus d'informations sur la manière dont cela peut améliorer les performances de votre charge de travail, consultez un rapport prix/performances 4,7 fois supérieur avec les instances AWS Graviton4 R8g utilisant Amazon Neptune v1.4.5.

Clusters de la bonne taille

Dimensionnez votre cluster en fonction de vos exigences en matière de simultanéité et de débit. Le nombre de requêtes simultanées pouvant être traitées par chaque instance du cluster est égal à deux fois le nombre de requêtes virtuelles CPUs (vCPUs) sur cette instance. Les requêtes supplémentaires qui arrivent alors que tous les threads de travail sont occupés sont placées dans une file d'attente côté serveur. Ces requêtes sont traitées sur une base first-in-first-out (FIFO) lorsque les threads de travail sont disponibles. La CloudWatch métrique `MainRequestQueuePendingRequests` Amazon indique la profondeur de file d'attente actuelle pour chaque instance. Si cette valeur est souvent supérieure à zéro, envisagez de choisir une instance avec plus de CPUs v. Si la profondeur de la file d'attente dépasse 8 192, Neptune renvoie `ThrottlingException` un message d'erreur.

Environ 65 % de la RAM de chaque instance est réservée au cache tampon. Le cache tampon contient l'ensemble de données de travail (pas l'intégralité du graphique, mais uniquement les données demandées). Pour déterminer le pourcentage de données extraites du cache tampon au lieu du stockage, surveillez la CloudWatch métrique`BufferCacheHitRatio`. Si cette métrique tombe souvent en dessous de 99,9 %, pensez à essayer une instance avec plus de mémoire pour déterminer si elle réduit votre latence et vos I/O coûts.

Les répliques de lecture ne doivent pas nécessairement avoir la même taille que votre instance Writer. Cependant, des charges de travail d'écriture importantes peuvent entraîner le retard des répliques plus petites et leur redémarrage, car elles ne peuvent pas suivre le rythme de la réplication. Par conséquent, nous recommandons de créer des répliques égales ou supérieures à l'instance du rédacteur.

Lorsque vous utilisez l'auto-scaling pour vos répliques de lecture, n'oubliez pas que la mise en ligne d'une nouvelle réplique de lecture peut prendre jusqu'à 15 minutes. Lorsque le trafic client augmente rapidement mais de manière prévisible, envisagez d'utiliser un dimensionnement planifié pour augmenter le nombre minimum de répliques de lecture afin de tenir compte de ce temps d'initialisation.

Les instances sans serveur prennent en charge différents cas d'utilisation et charges de travail. Envisagez de surprovisionner des instances sans serveur dans les scénarios suivants :

- Votre charge de travail fluctue souvent au cours de la journée.
- Vous avez créé une nouvelle application et vous n'êtes pas certain de la taille de la charge de travail.

- Vous êtes en train de développer et de tester.

Il est important de noter que les instances sans serveur sont plus chères que les instances provisionnées équivalentes sur la base d'un dollar par Go de RAM. Chaque instance sans serveur comprend 2 Go de RAM ainsi que le vCPU et le réseau associés. Effectuez une analyse des coûts entre vos options pour éviter les factures surprises. En général, vous réaliserez des économies avec le mode sans serveur uniquement lorsque votre charge de travail n'est très lourde que quelques heures par jour et presque nulle le reste de la journée ou si votre charge de travail fluctue de manière significative au cours de la journée.

Utilisez le calculateur de prix Amazon Neptune pour vous aider à évaluer la configuration appropriée pour votre cluster en fonction de facteurs tels que les exigences queries-per-second (QPS).

Optimisez les écritures

Pour optimiser les écritures, tenez compte des points suivants :

- Le Neptune Bulk Loader est le moyen optimal pour charger initialement votre base de données ou pour l'ajouter à des données existantes. Le chargeur Neptune n'est pas transactionnel et ne peut pas supprimer de données. Ne l'utilisez donc pas si vous le souhaitez.
- Les mises à jour transactionnelles peuvent être effectuées à l'aide des langages de requête pris en charge. Pour optimiser les I/O opérations d'écriture, écrivez les données par lots de 50 à 100 objets par validation. Un objet est un nœud, une arête ou une propriété sur un nœud ou une arête dans LPG, ou un triple store ou un quad dans RDF.
- Toutes les opérations d'écriture transactionnelles de Neptune sont effectuées en un seul thread pour chaque connexion. Lorsque vous envoyez une grande quantité de données à Neptune, envisagez d'avoir plusieurs connexions parallèles qui écrivent chacune des données. Lorsque vous choisissez une instance provisionnée Neptune, la taille de l'instance est associée à un nombre de v. CPUs Neptune crée deux threads de base de données pour chaque vCPU de l'instance. Commencez donc par deux fois le nombre de v CPUs lorsque vous testez une parallélisation optimale. Les instances sans serveur redimensionnent le nombre de v CPUs à un taux d'environ un pour 4 NCUs.

> **Note**
>
> Cela ne s'applique pas à l'API de chargement en masse, mais uniquement aux connexions directes.

- Planifiez et gérez efficacement tous ConcurrentModificationExceptionsles processus d'écriture, même si une seule connexion écrit des données à la fois. Concevez vos clients de manière fiable lorsque cela `ConcurrentModificationExceptions` se produit.
- Si vous souhaitez supprimer toutes vos données, pensez à utiliser l'API de réinitialisation rapide au lieu d'émettre des requêtes de suppression simultanées. Ce dernier prendra beaucoup plus de temps et entraînera des I/O coûts substantiels par rapport au premier.
- Si vous souhaitez supprimer la plupart de vos données, pensez à exporter les données que vous souhaitez conserver en utilisant neptune-export pour les charger dans un nouveau cluster. Supprimez ensuite le cluster d'origine.

Pilier d’optimisation des coûts

Le pilier d'optimisation des coûts du AWS Well-Architected Framework vise à éviter les coûts inutiles. Les recommandations suivantes peuvent vous aider à respecter les principes de conception d'optimisation des coûts et les meilleures pratiques architecturales pour Amazon Neptune.

Le pilier de l'optimisation des coûts se concentre sur les domaines clés suivants :

- Comprendre les dépenses au fil du temps et contrôler l'allocation des fonds
- Sélection des ressources du type et de la quantité appropriés
- Évolutivité pour répondre aux besoins de l'entreprise sans trop dépenser

Comprendre les modèles d'utilisation et les services nécessaires

Neptune convient parfaitement à votre charge de travail si votre modèle de données possède une structure graphique perceptible et que vos requêtes doivent explorer les relations et effectuer plusieurs sauts. Une base de données de graphes ne convient pas aux modèles suivants :

- Principalement des requêtes à saut unique (demandez-vous si vos données ne seraient pas mieux représentées sous forme d'attributs d'un objet)
- Données JSON ou BLOB stockées sous forme de propriétés
- Requêtes agrégées dans un ensemble de données, telles que le calcul de la somme d'une propriété numérique sur un grand nombre de nœuds

Déterminez si l'utilisation conjointe de plusieurs bases de données spécialement conçues pour des modèles d'accès spécifiques peut répondre à tous vos besoins. Par exemple :

- Une API qui nécessite des navigations graphiques complexes moins fréquentes ainsi que la récupération simultanée de propriétés pour un seul nœud peut être mieux présentée en utilisant une ou plusieurs options parmi Neptune, DynamoDB ou Amazon DocumentDB.
- Les bases de données relationnelles peuvent coexister avec Neptune pour conserver les fonctionnalités existantes, mais utilisez Neptune uniquement pour les traversées à sauts multiples qui ne fonctionnent pas et ne s'adaptent pas correctement dans les bases de données relationnelles.

Comprenez les coûts associés aux services qui interagissent avec Neptune et le complètent, notamment les suivants :

- Coûts de stockage d'Amazon Simple Storage Service (Amazon S3) pour les fichiers de données chargés en masse dans Neptune
- Fonctions Lambda utilisées pour les requêtes d'insertion ou d'insertion, les requêtes de lecture et le traitement des flux Neptune
- La couche API basée sur Neptune pour interagir avec l'application client (au lieu d'avoir des connexions directes à la base de données) dans Amazon API Gateway ou AWS AppSync
- AWS Glue tâches utilisées pour transférer des données vers et depuis Neptune
- Des instances Amazon Kinesis ou Amazon Managed Streaming for Apache Kafka (Amazon MSK) reçoivent des données de streaming pour une ingestion en temps quasi réel dans Neptune.
- AWS Database Migration Service pour la migration de données relationnelles vers Neptune
- Coûts SageMaker d'exécution d'Amazon pour les ordinateurs portables Jupyter et les modèles d'apprentissage automatique de la bibliothèque de graphes approfondis

Sélectionnez les ressources en tenant compte des coûts

La tarification de Neptune est basée sur le coût horaire de l'instance (ou les unités de calcul Neptune consommées en mode sans serveur), les E/S de données et l'utilisation du stockage. Les instances représentent, en moyenne, 85 % du coût total, de sorte que le bon dimensionnement peut avoir des implications financières importantes. Le meilleur moyen de dimensionner correctement les instances consiste à tester les performances des applications sur diverses instances et à comparer les facteurs suivants :

- La `MainRequestQueuePendingRequests` CloudWatch métrique reste-t-elle constamment à un chiffre bas proche de zéro ?
- L'`BufferCacheHitRatio` CloudWatch indicateur reste-t-il égal ou supérieur à 99,9 % la plupart du temps ?
- Quelles sont les courbes de coût et de performance, par exemple les coûts et les I/O coûts de données associés ? Les coûts de lecture des données peuvent augmenter de manière significative si une instance trop petite nécessite un échange fréquent entre le cache tampon et le stockage. `BufferCacheHitRatio`baissera fréquemment dans ces scénarios.

Les coûts des instances évoluent de manière linéaire en fonction de la taille au sein d'une même famille d'instances. Le coût horaire de l'`db.r6i.2xlarge`instance est le double de celui de l'`db.r6i.xlarge`instance et les ressources allouées sont également deux fois plus élevées. Le coût horaire de l'`db.r6i.24xlarge`instance est 24 fois supérieur à celui de l'`db.r6i.xlarge`instance.

Estimez le nombre de requêtes simultanées que vous devez prendre en charge. Vous pouvez avoir entre zéro et quinze répliques en lecture pour traiter les requêtes en lecture seule. Si vos besoins varient en fonction de l'heure de la journée, de la semaine ou du mois, vous pouvez utiliser plusieurs instances plus petites pour effectuer une mise à l'échelle selon un calendrier. Chaque vCPU d'une instance fournit deux threads pour gérer les requêtes simultanées. Trois répliques de `db.r6i.xlarge` lecture, avec 4 vCPU chacune, peuvent gérer 24 requêtes simultanées.

Si votre volume de trafic est plutôt mesuré en requêtes par seconde (QPS), vous devez expérimenter pour déterminer la latence moyenne de vos requêtes. Le nombre de requêtes par seconde qu'un cluster Neptune peut prendre en charge est égal à. `vCPU × 2 × (1 second/average query latency)` Par exemple, si vous avez 4 vCPU et une latence de requête de 100 millisecondes (0,1 seconde),. `QPS = 4 × 2 × (1s/0.1s) = 80 queries per second`

Les instances provisionnées sont moins chères que les instances sans serveur pour des charges de travail continues, stables et prévisibles. Le mode Serverless permet d'optimiser les coûts lorsque votre charge de travail nécessite une utilisation très élevée quelques heures par jour (par exemple`db.r6i.4xlarge`), puis pratiquement aucun trafic le reste de la journée (par exemple, une unité de calcul Neptune). Une instance sans serveur qui évolue pendant quelques heures puis redémarre sera moins coûteuse que d'utiliser une `db.r6i.4xlarge` instance provisionnée toute la journée.

Envisagez de passer à Neptune 1.4.5.0 ou version ultérieure et d'utiliser des `r8g` instances pour obtenir un meilleur débit de lecture et d'écriture à moindre coût que les instances d'ancienne génération, telles que ou. `r7g` `r6g` Pour plus d'informations, découvrez un rapport prix/performances de requête en écriture 4,7 fois supérieur avec les instances AWS Graviton4 R8g utilisant Amazon Neptune v1.4.5 (article de blog).AWS

Les clusters Neptune sont créés par défaut avec un stockage standard (si vous créez à l'aide de la console, le stockage est sélectionné par défaut, I/O-optimized storage). With I/O-optimized storage, you pay a slightly higher cost for storage and instances, but there are no I/O costs. This leads to more predictable recurring costs, but if your I/O usage is generally low, it may be more cost efficient to utilize standard storage. If you intend to load a lot of data initially, you can optimize cost by choosing I/O le chargement initial des données est effectué, puis il passe au stockage standard). Le type de stockage affecte uniquement le modèle de facturation et ne présente aucune différence technique

dans la configuration du cluster ou de l'instance de base de données Neptune. Vous pouvez modifier le type de stockage une fois tous les 30 jours. Au bout de 30 jours, vérifiez le détail de vos coûts Neptune et utilisez la page de tarification de Neptune pour déterminer si vos coûts auraient été plus élevés avec -optimized. I/O-optimized storage. If they would have been, continue to use standard storage, otherwise switch back to I/O

Choisissez la configuration d'instance Neptune la mieux adaptée à votre charge de travail

Si vous avez créé le vôtre Compte AWS avant le 15 juillet 2025, vous pouvez utiliser le niveau AWS gratuit pour des expériences d'entrée de gamme avec Neptune. Les 750 heures gratuites `db.t3.medium` et d'utilisation de l'`db.t4g.medium`instance sont suffisantes pour vous faire une bonne idée de Neptune à petite échelle. Votre cluster sera conservé après la fin de la période d'essai gratuite, mais son utilisation vous sera facturée à partir de cette date.

Les `db.t4g.medium` instances `db.t3.medium` et conviennent aux environnements de développement à faible coût dans lesquels vous n'utilisez pas OpenCypher, Graph Explorer ou diverses intégrations génératives d'IA. Ces instances ont un RAM-to-vCPU ratio inférieur (2:1) à celui des instances `R` familiales (8:1) ou des instances `X` familiales (16:1). Ce ratio réduit empêche l'utilisation des statistiques du moteur DFE qui permettent les performances d'OpenCypher, les intégrations GenAI (pour informer le LLM du schéma du graphe) et Graph Explorer. Les profils de performance peuvent être très différents lors de l'utilisation d'instances T familiales, en particulier pour les charges de travail mentionnées précédemment. Ces instances peuvent également augmenter le nombre de `OutOfMemoryExceptions` requêtes qui parcourent une partie significative du graphique. Pour déterminer si cette dernière condition est susceptible d'être affectée, vérifiez la `BufferCacheHitRatio` CloudWatch métrique.

Nous vous déconseillons vivement d'effectuer des tests de performance ou de charge avec des instances de T la famille, car vous pourriez obtenir des résultats incohérents qui ne sont pas révélateurs d'un environnement de production.

Les instances provisionnées vous offrent la meilleure combinaison de coûts et de performances lorsque votre charge de travail est relativement stable et prévisible. Choisissez la taille de l'instance en fonction de la simultanéité des demandes requise et de la complexité de la requête. Une simultanéité plus élevée nécessite plus de v. CPUs Une complexité de requête plus élevée nécessite davantage de RAM. Utilisez la `MainRequestQueuePendingRequests` CloudWatch métrique pour déterminer l'impact de la première option (une valeur supérieure à zéro représente un nombre

de demandes simultanées supérieur à ce qui peut être traité). Utilisez la `BufferCacheHitRatio` CloudWatch métrique pour déterminer l'impact de cette dernière. Un ratio souvent inférieur à 99,9 % indique qu'il n'y a pas assez de RAM pour contenir la partie active du graphique en cours d'évaluation, ce qui entraîne des échanges de cache plus fréquents. Si la famille d'instances R fournit une simultanéité suffisante mais pas assez de RAM, envisagez d'essayer la X famille d'instances.

Les cas d'utilisation idéaux pour les instances sans serveur sont décrits dans la documentation Neptune. Si vous ne savez pas si la solution provisionnée ou sans serveur est la meilleure solution pour vous et que le coût est votre principale préoccupation, testez votre charge de travail en mode sans serveur pour déterminer le nombre de machines NCUs utilisées et comparez le coût de provisioned (`N hours × hourly provisioned cost`) à celui de sans serveur (). `sum of NCUs × hourly cost per NCU` Si vous n'êtes pas sûr d'une instance de provisionnement de taille équivalente, une NCU équivaut à environ 2 Go de RAM, ainsi qu'au vCPU et au réseau associés. Si votre instance provisionnée appartient à la `r6i` famille, le ratio est de 1 vCPU pour 8 Go de RAM, ou NCUs 4, avec le réseau associé. Le calculateur de prix Amazon Neptune fournit également une comparaison pour vous aider à déterminer votre configuration de coûts optimale.

Lorsque vous utilisez le mode sans serveur pour les instances principales et de réplication, n'oubliez pas que les répliques en lecture des niveaux de promotion 0 et 1 seront redimensionnées NCUs en fonction de l'instance d'écriture afin qu'elles soient correctement dimensionnées en cas de basculement. Définissez vos limites de NCU pour ces instances en fonction des instances (enregistreurs ou lecteurs) qui reçoivent le plus de trafic.

Dans les environnements où le cluster n'est pas nécessaire 24 heures sur 24, 7 jours sur 7, envisagez d'écrire des scripts qui désactiveront les instances Neptune lorsqu'elles ne sont pas utilisées et les redémarreront avant leur utilisation. Les instances Neptune redémarrent automatiquement tous les 7 jours pour garantir l'application des mises à jour de maintenance requises. Si vous avez l'intention de laisser les instances désactivées pendant de longues durées, utilisez un script hebdomadaire pour les arrêter à nouveau.

Stockage et transfert de données à la bonne taille

Les requêtes plus efficaces (par exemple, les requêtes qui doivent toucher un nombre réduit de nœuds, d'arêtes et de propriétés dans le graphe) nécessitent moins de I/O transferts et peuvent éventuellement utiliser des instances plus petites car la quantité de cache tampon requise est moindre. Utilisez le profil ou expliquez les points de terminaison de votre langage de requête afin d'optimiser votre requête, et envisagez d'optimiser votre modèle de graphe en fonction des performances de vos requêtes.

Neptune utilise le codage par dictionnaire sur de grandes chaînes, et ce dictionnaire est optimisé pour les performances, et non pour l'efficacité. Si vous avez de grandes BLOBs chaînes JSON ou si vous changez fréquemment de chaîne pour les propriétés, envisagez de les stocker en dehors de Neptune dans Amazon S3, Amazon DynamoDB ou Amazon DocumentDB, et de stocker uniquement une référence dans le nœud Neptune.

Dans certains cas, le choix d'une taille d'instance plus grande peut s'avérer moins coûteux. Si vos I/O coûts sont très élevés en raison d'un faible niveau`BufferCacheHitRatio`, il est possible qu'un cache tampon plus important réduise considérablement ces coûts. En effet, toutes les données rentreraient dans le cache au lieu d'être fréquemment échangées depuis le stockage et d'augmenter le I/O taux de transfert.

Neptune utilise copy-on-write le clonage. Lorsque vous clonez pour diviser un graphe en plusieurs fragments, il peut être plus efficace de ne pas supprimer les données indésirables du cluster cloné, car cela impliquera la création de nouvelles pages de données, ce qui entraînera une augmentation des coûts de stockage. Les données inchangées par rapport à avant l'événement de clonage existeront sur une seule page de données partagée entre les deux clusters et ne seront facturées que pour cette copie unique.

N'activez pas l'index OSGP et n'utilisez pas d'instances R5d à moins d'avoir effectué des tests pour confirmer qu'elles font une différence substantielle dans votre charge de travail. Les deux sont conçus pour des scénarios rares, et ils peuvent augmenter vos coûts pour des gains minimes ou nuls.

Pilier de durabilité

Le pilier du développement durable vise à minimiser les impacts environnementaux liés à l'exécution de charges de travail dans le cloud. Les sujets clés incluent un modèle de responsabilité partagée pour la durabilité, la compréhension de l'impact et l'optimisation de l'utilisation afin de minimiser les ressources requises et de réduire les impacts en aval.

Le pilier du développement durable contient les principaux domaines d'intérêt suivants :

- Votre impact
- Objectifs de durabilité
- Utilisation maximisée
- Anticiper et adopter de nouvelles offres matérielles et logicielles plus efficaces
- Utilisation de services gérés
- Réduction de l'impact en aval

Ce guide met l'accent sur votre impact. Pour plus d'informations sur les autres principes de conception durable, consultez le AWS Well-Architected Framework.

Vos choix et vos exigences ont un impact sur l'environnement. Si vous pouvez choisir Régions AWS une solution à faible intensité en carbone et si vos exigences reflètent les besoins réels de la charge de travail au lieu de simplement maximiser le temps de disponibilité et la durabilité, la durabilité de la charge de travail augmente. Les sections suivantes traitent des meilleures pratiques et des considérations réfléchies qui auront un impact environnemental positif si elles sont adoptées dans la conception de votre charge de travail et dans les opérations en cours.

Région AWS sélection

Certains Régions AWS se trouvent à proximité de projets d'énergie renouvelable d'Amazon ou sont situés là où le réseau affiche une intensité en carbone publiée inférieure à celle d'autres. Tenez compte de l'impact sur le développement durable des régions qui pourraient être viables pour votre charge de travail et recoupez votre liste avec les régions dans lesquelles Neptune est disponible.

Consommation basée sur le comportement des utilisateurs

Le fait de bien dimensionner votre consommation en fonction du trafic et du comportement de vos utilisateurs permet de AWS minimiser l'impact des services sur l'environnement. Tenez compte des meilleures pratiques suivantes lors de la conception de votre solution :

- Surveillez CloudWatch les indicateurs Amazon tels que `CPUUtilization` `MainRequestQueuePendingRequests`, et `TotalRequestsPerSec` pour déterminer quand votre demande est la plus élevée et la plus faible, et assurez-vous que les ressources de votre cluster sont correctement dimensionnées pendant ces périodes.
- Automatisez l'arrêt des environnements hors production pendant les heures où ils ne sont pas utilisés. Pour plus d'informations, consultez le billet de blog Automatisez l'arrêt et le démarrage des ressources de l'environnement Amazon Neptune à l'aide de balises de ressources.
- Si vos modèles de trafic varient fréquemment et de manière imprévisible, pensez à utiliser des instances Neptune Serverless qui augmenteront ou diminueront en fonction de la demande au lieu d'utiliser une instance provisionnée pour les pics de trafic.
- Envisagez d'aligner vos accords de niveau de service sur les objectifs de durabilité en plus des objectifs de continuité des activités. L'assouplissement des exigences telles que la reprise après sinistre multirégionale, la haute disponibilité ou la conservation des sauvegardes à long terme, en particulier pour les environnements hors production ou les charges de travail non critiques, peut réduire la quantité de ressources nécessaires pour atteindre ces objectifs.

Optimisez le développement logiciel et les modèles d'architecture

Pour éviter le gaspillage, optimisez vos modèles et requêtes, et partagez les ressources de calcul afin d'utiliser toutes les ressources disponibles dans les instances et les clusters Neptune. Les meilleures pratiques spécifiques incluent :

- Demandez aux développeurs de partager les instances Neptune et les instances de l'application Jupyter Notebook au lieu de créer les leurs. Donnez à chaque développeur sa propre partition logique dans un seul cluster Neptune grâce à des stratégies de partitionnement multi-tenancy, et créez des dossiers de bloc-notes distincts pour chaque développeur sur une seule instance Jupyter.
- Mettez en œuvre des modèles qui optimisent l'utilisation des ressources et minimisent les temps d'inactivité, tels que des threads parallèles pour charger des données et regrouper des enregistrements dans le cadre d'une transaction plus importante.

- Optimisez vos requêtes et votre modèle graphique afin de minimiser les ressources nécessaires au calcul des résultats.
- Pour les résultats des requêtes Gremlin, utilisez la fonctionnalité de cache des résultats afin de minimiser les ressources dépensées pour recalculer les requêtes paginées ou récurrentes.
- Maintenez vos environnements Neptune à jour. Les dernières versions de Neptune prennent en charge les dernières instances Amazon EC2, telles que Graviton, qui sont plus efficaces. Ils proposent également des améliorations en matière d'optimisation des requêtes et des corrections de bogues qui réduisent la quantité de ressources nécessaires au calcul de vos requêtes.

Ressources

Références

- AWS Well-Architected
- AWS Documentation du framework Well-Architected
- Dernières mises à jour de Neptune
- Meilleures pratiques : tirer le meilleur parti de Neptune
- Calculateur de prix Amazon Neptune

Billets de blogs

- Tests automatisés de l'accès aux données Amazon Neptune avec Apache Gremlin TinkerPop
- Automatisez l'arrêt et le démarrage des ressources de l'environnement Amazon Neptune à l'aide de balises de ressources
- Contrôle d'accès précis pour les actions du plan de données Amazon Neptune
- Rapport prix/performances de requête en écriture 4,7 fois supérieur avec les instances AWS Graviton4 R8g utilisant Amazon Neptune v1.4.5
- Comment Orca Security a optimisé les performances de sa base de données Amazon Neptune
- Créez des applications graphiques plus rapidement avec les points de terminaison publics Amazon Neptune
- La nouvelle version du moteur Amazon Neptune fournit un débit jusqu'à 9 fois plus rapide et 10 fois supérieur pour les performances des requêtes OpenCypher

Cours gratuits AWS de création de compétences

- Commencer à utiliser Amazon Neptune
- Création d'applications sur Amazon Neptune
- Modélisation des données pour Amazon Neptune

Collaborateurs

Les contributeurs à ce guide incluent :

- Brian O'Keefe, architecte principal des solutions Neptune, AWS
- Abhishek Mishra, architecte senior des solutions Neptune, AWS
- Ganesh Sawhney, chef d'équipe - Architecte de solutions de réussite pour les partenaires stratégiques, AWS
- Michael Havey, architecte principal des solutions Neptune, AWS
- Kevin Phillips, architecte des solutions Neptune, AWS
- Melissa Kwok, architecte des solutions Neptune, AWS
- Sakti Mishra, architecte de solutions principal AWS
- Javed Ali, architecte de solutions senior, AWS

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 de Neptune	Nous avons mis à jour la documentation pour inclure des informations sur Amazon Neptune 1.4.6.0 et versions ultérieures.	2 janvier 2026
Publication initiale	—	27 septembre 2023

AWS Glossaire des directives prescriptives

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

Nombres

7 R

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

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

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

A

ABAC

Voir contrôle d'accès basé sur les attributs.

services abstraits

Consultez la section Services gérés.

ACIDE

Voir atomicité, consistance, isolation, durabilité.

migration active-active

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

migration active-passive

Méthode de migration de base de données dans laquelle 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 Gestion des identités et des accès AWS (IAM).

source de données faisant autorité

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

Zone de disponibilité

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

AWS Cadre d'adoption du cloud (AWS CAF)

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

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

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

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

B

mauvais bot

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

BCP

Consultez la section Planification de la continuité des activités.

graphique de comportement

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

système de poids fort

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

classification binaire

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

filtre de Bloom

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

déploiement bleu/vert

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

bot

Application logicielle qui exécute des tâches automatisées sur Internet et simule l'activité ou l'interaction humaine. Certains robots sont utiles ou bénéfiques, comme les robots d'exploration Web qui indexent des informations sur Internet. D'autres robots, 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 Gestion des identités et des accès AWS (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 CSPM 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 cadre AWS Well-Architected.

compte membre

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

MAILLES

Voir le système d'exécution de la fabrication.

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

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

microservice

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 des responsables des opérations, des analystes commerciaux et des propriétaires, des ingénieurs de migration, des développeurs et DevOps des professionnels travaillant dans le cadre de sprints.

Entre 20 et 50 % du portefeuille d'applications d'entreprise est constitué de modèles répétés qui peuvent être optimisés par une approche d'usine. Pour plus d'informations, veuillez consulter la rubrique discussion of migration factories et le guide Cloud Migration Factory dans cet ensemble de contenus.

métadonnées de migration

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

modèle de migration

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

Évaluation du portefeuille de migration (MPA)

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

Évaluation de la préparation à la migration (MRA)

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

stratégie de migration

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

ML

Voir apprentissage automatique.

modernisation

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

évaluation de la préparation à la modernisation

Évaluation qui permet de déterminer si les applications d'une organisation sont prêtes à être modernisées, d'identifier les avantages, les risques et les dépendances, et qui détermine dans quelle mesure l'organisation peut prendre en charge l'état futur de ces applications. Le résultat de l'évaluation est un plan de l'architecture cible, une feuille de route détaillant les phases de développement et les étapes du processus de modernisation, ainsi qu'un plan d'action pour combler les lacunes identifiées. Pour plus d'informations, consultez la section Évaluation de l'état de préparation à la modernisation des applications dans le AWS Cloud.

applications monolithiques (monolithes)

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

MPA

Voir Évaluation du portefeuille de migration.

MQTT

Voir Message Queuing Telemetry Transport.

classification multi-classes

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

infrastructure mutable

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

O

OAC

Voir Contrôle d'accès à l'origine.

OAI

Voir l'identité d'accès à l'origine.

OCM

Voir gestion du changement organisationnel.

migration hors ligne

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

OI

Consultez la section Intégration des opérations.

OLA

Voir l'accord au niveau opérationnel.

migration en ligne

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

OPC-UA

Voir Open Process Communications - Architecture unifiée.

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

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

accord au niveau opérationnel (OLA)

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

examen de l'état de préparation opérationnelle (ORR)

Une liste de questions et de bonnes pratiques associées qui vous aident à comprendre, 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 peuvent accéder au contenu d'un compartiment S3 uniquement via 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.

policy

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.

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 les 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).

RAG

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 en matière AWS de sécurité. Parmi les actions de réponse automatique, citons la modification d'un groupe de sécurité VPC, l'application de correctifs à une instance Amazon EC2 ou la rotation des informations d'identification.

chiffrement côté serveur

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

Politique de contrôle des services (SCP)

Politique qui 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, voir 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

tags

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.