Appliquer le framework AWS Well-Architected pour Amazon Neptune Analytics

AWS Conseils prescriptifs

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

Table of Contents

Appliquer le framework AWS Well-Architected pour Amazon Neptune Analytics

Michael Havey, Amazon Web Services ()AWS

Décembre 2024 (historique du document)

Vous pouvez créer des solutions basées sur des graphiques sur Amazon Web Services (AWS) à l'aide d'Amazon Neptune. Neptune inclut Neptune Analytics, un moteur d'analyse graphique optimisé pour la mémoire qui peut analyser rapidement de grandes quantités de données graphiques pour obtenir des informations et identifier des tendances. Il peut effectuer des analyses sur les données de votre cluster de base de données Neptune existant, ou vous pouvez charger et analyser des données provenant d'ensembles de données externes. Ce guide fournit des conseils prescriptifs pour appliquer les principes du AWS Well-Architected Framework lorsque vous planifiez votre déploiement de Neptune Analytics. L'application du AWS framework Well-Architected pour Amazon Neptune couvre le même sujet pour une base de données 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 piliers de conception et les meilleures pratiques du Well-Architected Framework, ainsi que les points à prendre en compte lorsque vous déployez Neptune Analytics 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 prises en charge.
- Suivez les modèles de conception de AWS Well-Architected qui vous aident 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 graphe Neptune Analytics 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 et en utilisant les métriques et les mécanismes d'Amazon Neptune Analytics 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 à l'aide d'une approche iAC

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

- Appliquez IaC pour déployer les graphiques de Neptune Analytics et les ressources associées. Pour une configuration cohérente de l'environnement, utilisez le support de Neptune Analytics fourni par pour fournir des graphiques et AWS CloudFormationdes points de terminaison privés.
- CloudFormation À utiliser pour approvisionner des instances de blocs-notes Neptune sur Amazon SageMaker AI. Vous pouvez utiliser des blocs-notes pour interroger et visualiser des données dans un graphe Neptune Analytics.

- Lorsque vous créez un graphe Neptune Analytics à partir d'une source existante, telle qu'un cluster de base de données ou un instantané Neptune, ou des fichiers de données stockés dans Amazon Simple Storage Service (Amazon S3), surveillez la tâche d'importation en masse.
- Automatisez les procédures opérationnelles de Neptune Analytics, telles que le redimensionnement du graphique, sa suppression et sa capture instantanée, la restauration du graphique à partir d'un instantané, ainsi que la réinitialisation et le rechargement du graphique. Utilisez l'API Neptune Analytics, disponible via le AWS Command Line Interface (AWS CLI) ou les SDK.
- Évaluez le temps de disponibilité requis pour votre graphe. Les analyses sont souvent éphémères ; le graphique n'est nécessaire que pendant le temps nécessaire à l'exécution des algorithmes. Si tel est le cas, utilisez le AWS CLI ou SDKs pour prendre un cliché et supprimer le graphique lorsqu'il n'est plus nécessaire. Vous pourrez ensuite le restaurer ultérieurement à partir d'un instantané, si nécessaire.
- Stockez les chaînes de connexion en dehors de votre client. Vous pouvez stocker les chaînes de connexion dans AWS Secrets ManagerAmazon DynamoDB ou dans n'importe quel endroit où elles peuvent être modifiées dynamiquement.
- Utilisez des balises pour ajouter des métadonnées à vos ressources Neptune Analytics et suivez l'utilisation en fonction des balises. Les balises vous aident à organiser vos ressources. Par exemple, vous pouvez appliquer une balise commune aux ressources d'un environnement ou d'une application spécifique. Vous pouvez également utiliser des balises pour analyser la facturation de l'utilisation des ressources ; pour plus d'informations, voir Organisation et suivi des coûts à l'aide des balises de répartition des AWS coûts dans le Guide AWS de l'utilisateur de facturation. En outre, vous pouvez utiliser des conditions dans vos politiques AWS Identity and Access Management (IAM) pour contrôler l'accès aux AWS ressources en fonction des balises utilisées sur ces ressources. Pour cela, vous devez utiliser la clé de condition globale `aws:ResourceTag/tag-key`. Pour plus d'informations, consultez la section Contrôle de l'accès aux AWS ressources dans le Guide de l'utilisateur IAM.

Conception pour les opérations

Adoptez des approches pour améliorer la façon dont vous exploitez les graphiques de Neptune Analytics :

- Conservez des graphiques Neptune Analytics distincts pour le développement, les tests et la production. Ces graphiques peuvent avoir des ensembles de données, des utilisateurs et des contrôles opérationnels différents.
- Conservez des graphiques Neptune Analytics distincts pour différentes utilisations. Par exemple, si deux groupes d'utilisateurs analytiques ont besoin de graphiques distincts avec des chronologies, des modèles, des performances, une disponibilité et des SLAs modèles d'utilisation différents, maintenez des graphiques distincts pour chaque groupe.
- Préparez les utilisateurs et le personnel opérationnel aux mises à jour de maintenance de Neptune Analytics.

Effectuez des modifications fréquentes, mineures et réversibles

Les recommandations suivantes se concentrent sur de petites modifications réversibles que vous pouvez apporter 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 CodeDeployAWS CodeBuild Compilez, testez et déployez du code dans un environnement Neptune Analytics hors production avant de le promouvoir en tant que graphe de production.

Mettez en œuvre l'observabilité pour obtenir des informations exploitables

Obtenez une compréhension complète du comportement, des performances, de la fiabilité, des coûts et de l'intégrité des charges de travail. Les recommandations suivantes vous aident à acquérir ce niveau de compréhension dans Neptune Analytics :

- Surveillez CloudWatch les métriques Amazon pour Neptune Analytics. À partir de ces indicateurs, vous pouvez déterminer la taille d'un graphe (nombre de nœuds, d'arêtes et de vecteurs, plus la taille totale en octets), l'utilisation du processeur et les taux d'erreur et de demandes de requêtes.
- Créez CloudWatch des tableaux de bord et des alarmes pour les indicateurs clés tels que `NumQueuedRequestsPerSec` `NumOpenCypherRequestsPerSecGraphStorageUsagePercent`,,`GraphSizeBytes`, et `CPUUtilization` les réponses des clients Neptune figurant dans les journaux de vos applications.
- Définissez des notifications pour surveiller l'état du graphique Neptune Analytics, par exemple lorsque la taille du graphique, le taux de demandes ou l'utilisation du processeur dépassent votre seuil. Par exemple, si la valeur `GraphStorageUsagePercent` est passée à 90 % sur un graphique que vous comptez augmenter de manière significative, décidez si vous souhaitez augmenter la capacité de l'unité de capacité Neptune (m-NCU) optimisée pour la mémoire. Si le m-NCU actuel est de 128, l'augmenter à 256 réduira le stockage d'environ 45 %. Si elle `NumQueuedRequestsPerSec` est souvent supérieure à zéro, envisagez d'augmenter la capacité m-NCU pour augmenter la capacité de calcul. Vous pouvez également réduire la simultanéité côté client.

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 vous le souhaiteriez. 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 de votre 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

Utilisez la journalisation pour observer les modèles de performance et d'activité anormaux. Tenez compte des meilleures pratiques suivantes :

- Neptune Analytics prend en charge la journalisation des actions du plan de contrôle en utilisant. AWS CloudTrail Pour plus d'informations, consultez la section Journalisation des appels d'API Neptune Analytics à l'aide de. AWS CloudTrail Grâce à cette fonctionnalité, vous pouvez suivre la création, la mise à jour et la suppression des ressources Neptune Analytics. Pour une surveillance et des alertes robustes, vous pouvez également intégrer CloudTrail des événements à Amazon CloudWatch Logs. Pour améliorer votre analyse de l'activité des services Neptune Analytics et identifier les modifications des activités d'un Compte AWS, vous pouvez interroger les CloudTrail journaux à l'aide d'Amazon Athena. Par exemple, vous pouvez utiliser des requêtes pour identifier des tendances et isoler davantage l'activité par attributs, comme l'adresse IP source ou l'utilisateur.
- Vous pouvez également l'utiliser CloudTrail pour activer la journalisation des activités du plan de données Neptune Analytics, telles que les exécutions de requêtes. Vous pouvez voir quelles requêtes sont exécutées, leur fréquence et leur source. Par défaut, CloudTrail n'enregistre pas les événements liés aux données. Des frais supplémentaires sont facturés pour les événements de données. Pour en savoir plus, consultez Tarification de AWS CloudTrail.
- Vous pouvez également enregistrer les appels de vos applications à Neptune Analytics dans le plan de contrôle ou dans le plan de données. Par exemple, si vous utilisez le AWS SDK pour Python (Boto3)pour effectuer des requêtes, vous pouvez activer la journalisation au niveau du débogage pour obtenir une trace des requêtes dans la console ou le fichier. Cela est utile pendant le développement. Nous vous recommandons également de intercepter et de consigner les exceptions issues de votre application.

Pilier de sécurité

La sécurité du cloud est la priorité absolue de AWS. 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 et AWS. Le modèle de responsabilité partagée décrit cette notion par les termes 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 la AWS sécurité dans le cadre des programmes de AWS conformité. Pour en savoir plus sur les programmes de conformité qui s'appliquent à Neptune, consultez la section AWS Services concernés par 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 la section Confidentialité des données FAQs. 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 du AWS Well-Architected Framework vous aide à comprendre comment appliquer le modèle de responsabilité partagée lorsque vous utilisez Neptune Analytics. Les rubriques suivantes expliquent comment configurer Neptune Analytics pour atteindre vos objectifs de sécurité et de conformité. Vous apprendrez également à utiliser d'autres outils Services AWS qui vous aident à surveiller et à sécuriser vos ressources Neptune Analytics. 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 graphes, les balises, les rôles 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.
- Un graphe Neptune Analytics est chiffré au repos. Vous pouvez utiliser la clé par défaut ou une clé AWS Key Management Service (AWS KMS) de votre choix pour chiffrer le graphe. Vous pouvez également chiffrer les instantanés et les données exportés vers Amazon S3 lors de l'importation en masse. Vous pouvez supprimer le chiffrement une fois l'importation terminée.
- Lorsque vous utilisez le langage OpenCypher, appliquez les 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 la section Exemples de requêtes paramétrées OpenCypher dans la documentation Neptune.

Sécurisez vos réseaux

Vous pouvez activer un graphe Neptune Analytics pour la connectivité publique afin qu'il soit accessible depuis l'extérieur d'un cloud privé virtuel (VPC). Cette connectivité est désactivée par défaut. Le graphique nécessite une authentification IAM. L'appelant doit obtenir une identité et être autorisé à utiliser le graphique. Par exemple, pour exécuter une requête OpenCypher, l'appelant doit disposer d'autorisations de lecture, d'écriture ou de suppression sur le graphe spécifique.

Vous pouvez également créer des points de terminaison privés pour le graphe afin d'accéder au graphe depuis un VPC. Lorsque vous créez le point de terminaison, vous spécifiez le VPC, les sous-réseaux et les groupes de sécurité pour restreindre l'accès à l'appel du graphe.

Pour protéger vos données en transit, Neptune Analytics applique des connexions SSL via HTTPS au graphe. Pour plus d'informations, consultez la section Protection des données dans Neptune Analytics dans la documentation de Neptune Analytics.

Mettre en œuvre l'authentification et l'autorisation

Les appels à un graphe Neptune Analytics nécessitent une authentification IAM. L'appelant doit obtenir une identité et disposer des autorisations suffisantes pour effectuer l'action sur le graphique. Pour une description des actions d'API et de leurs autorisations requises, consultez la documentation de l'API Neptune Analytics. Vous pouvez appliquer des contrôles d'état pour restreindre l'accès par balise.

L'authentification IAM utilise le protocole AWS Signature Version 4 (SigV4). Pour simplifier l'utilisation depuis votre application, nous vous recommandons d'utiliser un AWS SDK. Par exemple, en Python, utilisez le client Boto3 pour Neptune Graph, qui extrait SigV4.

Lorsque vous chargez des données dans le graphique, le chargement par lots utilise les informations d'identification IAM de l'appelant. L'appelant doit être autorisé à télécharger des données depuis Amazon S3 avec la relation de confiance configurée afin que Neptune Analytics puisse assumer le rôle de charger les données dans le graphique à partir des fichiers Amazon S3.

L'importation en bloc peut être effectuée soit lors de la création du graphe (par l'équipe d'infrastructure), soit sur un graphique vide existant (par l'équipe d'ingénierie des données autorisée à démarrer des tâches d'importation). Dans les deux cas, Neptune Analytics assume le rôle IAM fourni par l'appelant en entrée. Ce rôle lui donne l'autorisation de lire et de répertorier le contenu du dossier Amazon S3 dans lequel les données d'entrée sont stockées.

Pilier de fiabilité

Le AWS pilier de fiabilité du Well-Architected Framework englobe la capacité d'une charge de travail à exécuter correctement et de manière cohérente la fonction prévue, au moment prévu. Cela inclut la capacité 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 ont un impact sur le comportement de votre charge de travail dans tous les piliers de Well-Architected. Pour des raisons de fiabilité, il existe des modèles spécifiques que vous devez suivre, comme indiqué dans cette section.

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

Votre AWS compte dispose de quotas par défaut (anciennement appelés limites) pour chacun d'entre eux Service AWS. Sauf indication contraire, chaque quota est spécifique à la région. Vous pouvez demander des augmentations pour certains quotas, mais pas pour tous.

Pour trouver des quotas pour Neptune Analytics, ouvrez la console Service Quotas. Dans le volet de navigation, choisissez Services AWS, puis sélectionnez Amazon Neptune Analytics. Faites attention aux quotas relatifs au nombre de graphiques et d'instantanés, à la mémoire maximale allouée à un graphique et aux taux de demandes d'API.

Si la mémoire maximale allouée n'est pas suffisante pour votre ensemble de données, déterminez quels types de nœuds et de bords sont essentiels pour l'utilisation analytique que vous souhaitez en faire. Chargez un sous-ensemble de données afin que les analyses soient possibles dans les limites de la capacité allouée autorisée. De nombreuses charges de travail d'analyse, en particulier celles qui exécutent des algorithmes de graphes, nécessitent uniquement la topologie avec un ensemble limité de propriétés au lieu du graphe transactionnel complet. (Pour une discussion sur les différences entre les charges de travail transactionnelles et analytiques, voir la section sur le pilier de l'efficacité des performances.)

Si le nombre maximum de graphiques n'est pas suffisant pour l'usage que vous souhaitez en faire :

- Envisagez de combiner des graphiques ayant des utilisations similaires.
- Évaluez le nombre de graphiques qui doivent être exécutés à un moment donné. Si vous avez un cas d'utilisation éphémère de l'analytique, capturez un graphique et supprimez-le lorsqu'il n'est plus nécessaire. Cela réduit le nombre de graphiques par rapport au quota.
- Envisagez de créer des graphiques de provisionnement sous différentes formes. Comptes AWS

Comprendre les modèles de déploiement de Neptune

Prenez en compte les points de décision suivants lorsque vous envisagez de déployer un graphe Neptune Analytics :

- Ensemencement : décidez si vous souhaitez créer un graphique vide ou y charger des données au moment de sa création avec des données provenant d'Amazon S3, d'un cluster de base de données Neptune existant ou d'un instantané de base de données Neptune existant.

  Recommandation : Si la source est un cluster ou un instantané Neptune, vous devez charger ses données au moment de la création du graphe. Si la source est Amazon S3, chargez les données au moment de leur création si l'effort de chargement est important et qu'il est préférable de les exécuter dans le cadre d'une activité de provisionnement d'infrastructure. Si vous préférez charger des données dans le cadre d'une activité d'ingénierie des données ou d'application, créez un graphique vide et chargez les données depuis Amazon S3 ultérieurement.
- Capacité : estimez la capacité provisionnée requise pour un graphique, en fonction de la taille des données et de l'utilisation prévue des applications.

  Recommandation : Au moment de la création, spécifiez la mémoire maximale allouée afin de limiter la taille du graphique. Ce paramètre est obligatoire. Vous pouvez modifier la capacité ultérieurement si nécessaire.
- Disponibilité et tolérance aux pannes : déterminez si des répliques sont nécessaires pour garantir la disponibilité. Une réplique fait office de réserve pour la restauration en cas de défaillance du graphe. Un graphe contenant des répliques se rétablit plus rapidement qu'un graphe dépourvu de répliques. Déterminez également combien de temps le graphique est nécessaire, s'il est destiné à des analyses éphémères uniquement et, dans l'affirmative, à quel moment il sera supprimé.

Recommandation : déterminez les exigences de disponibilité, telles que la durée pendant laquelle le graphique peut être indisponible et le moment où il peut être supprimé, avant de créer un graphique.

- Mise en réseau et sécurité : déterminez si vous avez besoin d'une connectivité publique, d'une connectivité privée ou des deux, et si vous souhaitez chiffrer vos données.

  Recommandation : Avant de créer un graphique, déterminez les exigences de l'organisation, par exemple si la connectivité publique est autorisée et où les applications clientes seront déployées.
- Sauvegardes et restauration : déterminez si des instantanés doivent être créés et, dans l'affirmative, à quel moment et dans quelles conditions. Déterminez si votre organisation a des exigences en matière de reprise après sinistre (DR).

  Recommandation : La création d'instantanés est une activité manuelle. Décidez à quel moment créer des instantanés et prenez en compte vos exigences en matière de reprise après sinistre avant de créer un graphique.

Gérez et dimensionnez les clusters Neptune

Un graphe Neptune Analytics se compose d'une seule instance optimisée pour la mémoire. La capacité (m-NCU) de l'instance est définie au moment de sa création. L'instance peut être mise à l'échelle verticale en augmentant la capacité allouée par le biais d'une action administrative ; la capacité allouée peut également être réduite. Les répliques étant des cibles de basculement passives, elles n'augmentent pas l'échelle d'un graphique. À cet égard, une réplique de graphe est différente d'une réplique de lecture de base de données Neptune, qui est une instance active dans un cluster Neptune capable de traiter les opérations de lecture à partir d'applications.

Les répliques ont un coût. Le prix de la réplique est calculé au taux de m-ncu indiqué sur le graphique. Par exemple, si un graphe est provisionné pour 128 m-NCU et possède une seule réplique, le coût est deux fois supérieur à celui d'un graphe équivalent sans répliques.

Dans le domaine de l'analytique, il existe deux raisons principales de passer à l'échelle supérieure :

- Afin de fournir davantage de mémoire et de processeur pour les requêtes analytiques et les algorithmes, étant donné que chaque requête individuelle est coûteuse, l'algorithme graphique à exécuter est intrinsèquement complexe et nécessite davantage de ressources compte tenu de ses entrées, ou le taux de demandes simultanées est élevé. Si de telles requêtes rencontrent des out-of-memory erreurs, la mise à l'échelle est une solution raisonnable.

- Pour prendre en charge une taille de graphique supérieure à celle prévue. Par exemple, si la capacité actuellement allouée est de 128 M-NCU pour prendre en charge 60 Go de données source et que vous avez besoin de 40 Go supplémentaires de données sources, une augmentation à 256 M-NCU est justifiée.

Surveillez CloudWatch les métriques pour Neptune Analytics, telles que,`NumQueuedRequestsPerSec`,, et `NumOpenCypherRequestsPerSec` `GraphStorageUsagePercent` `GraphSizeBytesCPUUtilization`, afin de déterminer si une mise à l'échelle est nécessaire. Vous pouvez mettre à jour la configuration d'un graphique via la console AWS CLI, ou SDKs. (Pour des exemples et des meilleures pratiques, voir la section sur le pilier de l'excellence opérationnelle.)

Gestion des sauvegardes et des événements de basculement

Utilisez des répliques pour garantir la disponibilité d'un graphe en cas de panne. Un graphique utilise la persistance basée sur les journaux pour valider les modifications dans les zones de disponibilité d'un Région AWS. La réplique agit en mode veille et a accès à ces données. En cas de défaillance, le graphe reprend les opérations sur le réplica. L'application continue d'utiliser le même point de terminaison pour se connecter au graphe. Les demandes en vol pendant la panne génèrent des erreurs avec une exception d'indisponibilité du service. Envisagez d'utiliser un schéma de nouvelle tentative avec interruption dans le code de l'application pour détecter l'erreur, puis réessayez après un bref intervalle. Les nouvelles demandes effectuées pendant le basculement sont mises en file d'attente et peuvent connaître une latence plus longue.

Si aucune réplique n'est configurée et que le graphe échoue, Neptune Analytics se rétablit à partir d'un stockage durable, mais la restauration prend plus de temps car Neptune doit réinitialiser les ressources.

Créez des instantanés du graphique. (Neptune Analytics ne prend pas de clichés automatiques.) Si le graphique est régulièrement modifié après sa création, prenez fréquemment des instantanés pour capturer son état actuel. Supprimez les anciens instantanés s'il n'est pas nécessaire de les restaurer à une date antérieure.

Vous pouvez partager des instantanés avec d'autres comptes et entre eux. Régions AWS Si vous avez des exigences en matière de reprise après sinistre, déterminez si la restauration du graphe dans une autre région à partir d'un instantané répond à vos exigences en matière de temps de restauration (RTO) et d'objectif de point de restauration (RPO).

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
- Identifier les composants peu 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 vous 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 Neptune Analytics.

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 graphique
- Optimisation de l'écriture

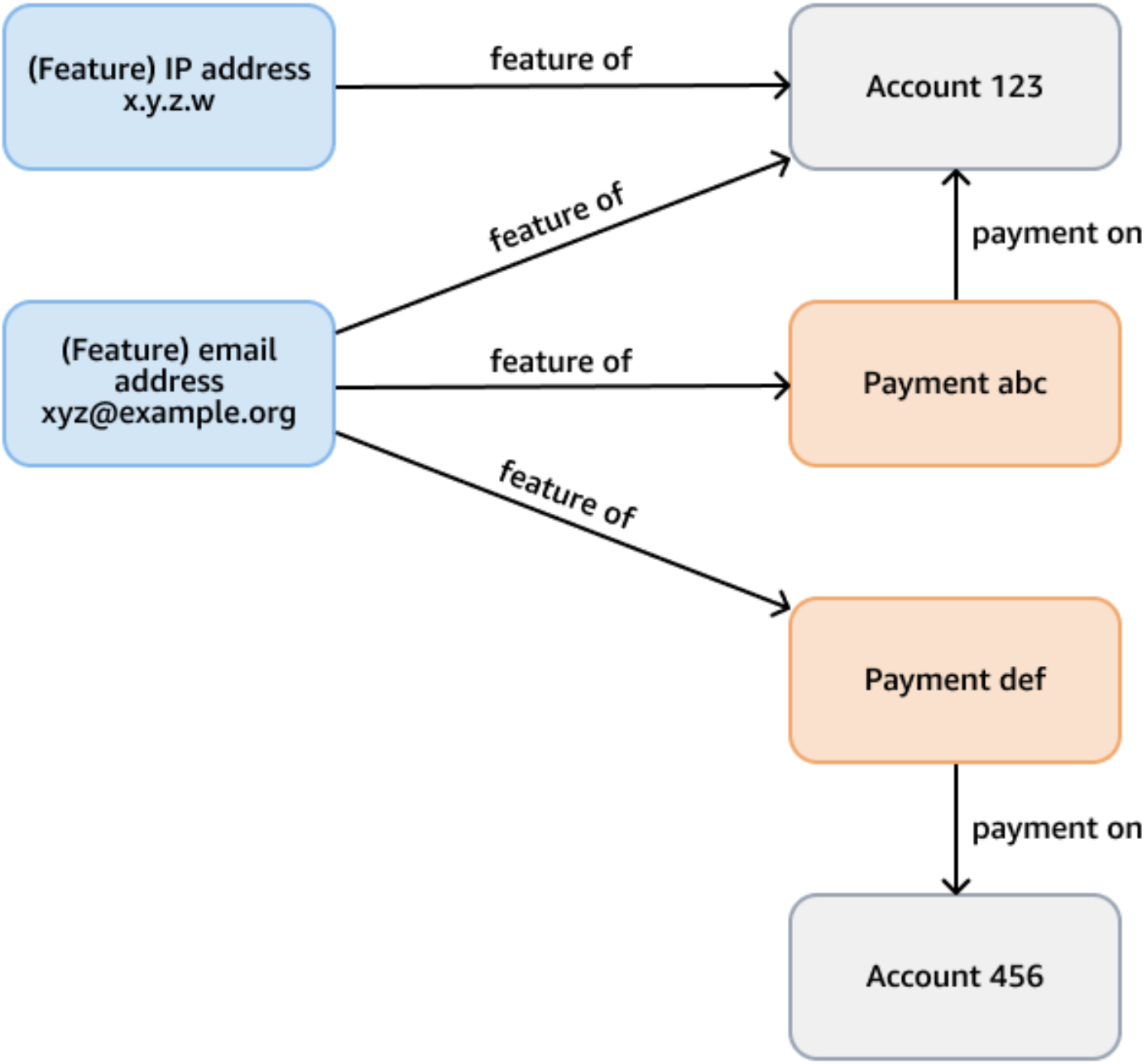
Comprendre la modélisation de graphes à des fins d'analyse

Le guide Applying the AWS Well-Architected Framework for Amazon Neptune traite de la modélisation de graphes pour améliorer les performances. Les décisions de modélisation qui affectent les performances incluent le choix des nœuds et des arêtes requis, de leurs IDs étiquettes et de leurs propriétés, de la direction des arêtes, du caractère générique ou spécifique des étiquettes et, d'une manière générale, de l'efficacité avec laquelle le moteur de requêtes peut naviguer dans le graphe pour traiter les requêtes courantes.

Ces considérations s'appliquent également à Neptune Analytics ; il est toutefois important de faire la distinction entre les modèles d'utilisation transactionnels et analytiques. Un modèle de graphe

efficace pour les requêtes dans une base de données transactionnelle telle qu'une base de données Neptune devra peut-être être remodelé à des fins d'analyse.

Prenons l'exemple d'un graphique de fraude dans une base de données Neptune dont le but est de détecter les tendances frauduleuses dans les paiements par carte de crédit. Ce graphique peut comporter des nœuds représentant les comptes, les paiements et les fonctionnalités (telles que l'adresse e-mail, l'adresse IP, le numéro de téléphone) du compte et du paiement. Ce graphe connecté prend en charge des requêtes telles que le fait de parcourir un chemin de longueur variable qui part d'un paiement donné et nécessite plusieurs sauts pour trouver les fonctionnalités et les comptes associés. La figure suivante montre un tel graphique.



L'exigence d'analyse peut être plus spécifique, par exemple pour trouver des communautés de comptes liées par une fonctionnalité. Vous pouvez utiliser l'algorithme WCC (Weakly Connected Components) à cette fin. Il est inefficace de l'exécuter par rapport au modèle de l'exemple précédent, car il doit traverser plusieurs types de nœuds et d'arêtes. Le modèle du schéma suivant est plus efficace. Il relie les `account` nœuds à un `shares feature` avantage si les comptes eux-mêmes (ou les paiements provenant des comptes) partagent une même fonctionnalité. Par exemple, `Account 123` possède une fonctionnalité `xyz@example.org` de courrier électronique et `Account 456` utilise le même e-mail pour un paiement (`Payment def`).

La complexité informatique du WCC `|E|` est `O(|E|logD)` le nombre d'arêtes dans le graphique et `D` le diamètre (la longueur du chemin le plus long) qui relie les nœuds. Comme le modèle transactionnel omet les nœuds non essentiels, il optimise à la fois le nombre d'arêtes et le diamètre, et réduit la complexité de l'algorithme WCC.

Lorsque vous utilisez Neptune Analytics, revenez à partir des algorithmes et des requêtes analytiques requis. Si nécessaire, modifiez le modèle pour optimiser ces requêtes. Vous pouvez remodeler le modèle avant de charger des données dans le graphique ou d'écrire des requêtes qui modifient les données existantes dans le graphique.

Optimisation des requêtes

Suivez ces recommandations pour optimiser les requêtes Neptune Analytics :

- Utilisez les requêtes paramétrées et le cache du plan de requêtes, qui est activé par défaut. Lorsque vous utilisez le cache du plan, le moteur prépare la requête pour une utilisation ultérieure, à condition qu'elle soit terminée en 100 millisecondes ou moins, ce qui permet de gagner du temps lors des appels suivants.
- Pour les requêtes lentes, exécutez un plan d'explication pour identifier les goulots d'étranglement et apporter des améliorations en conséquence.
- Si vous utilisez la recherche par similarité vectorielle, déterminez si des incorporations plus petites produisent des résultats de similarité précis. Vous pouvez créer, stocker et rechercher des intégrations plus petites de manière plus efficace.
- Suivez les meilleures pratiques documentées pour utiliser OpenCypher dans Neptune Analytics. Par exemple, utilisez des cartes aplaties dans une clause UNWIND et spécifiez des étiquettes de bord dans la mesure du possible.
- Lorsque vous utilisez un algorithme graphique, comprenez les entrées et les sorties de l'algorithme, sa complexité de calcul et, d'une manière générale, son fonctionnement.

- Avant d'appeler un algorithme graphique, utilisez une `MATCH` clause pour minimiser l'ensemble de nœuds en entrée. Par exemple, pour limiter les nœuds à partir desquels effectuer une recherche basée sur l'étendue (BFS), suivez les exemples fournis dans la documentation de Neptune Analytics.
- Filtrez sur les étiquettes des nœuds et des arêtes si possible. Par exemple, BFS dispose de paramètres d'entrée pour filtrer le passage vers une étiquette de nœud spécifique (`vertexLabel`) ou des étiquettes de bord spécifiques (`edgeLabels`).
- Utilisez des paramètres de délimitation, par exemple `maxDepth` pour limiter les résultats.
- Expérimentez avec le `concurrency` paramètre. Essayez-le avec une valeur de 0, qui utilise tous les threads d'algorithme disponibles pour paralléliser le traitement. Comparez cela à l'exécution monothread en définissant le paramètre sur 1. Un algorithme peut être exécuté plus rapidement dans un seul thread, en particulier sur des entrées plus petites, telles que les recherches de faible envergure, où le parallélisme n'entraîne aucune réduction mesurable du temps d'exécution et peut entraîner une surcharge.
- Choisissez entre des types d'algorithmes similaires. Par exemple, Bellman-Ford et Delta-Stepping sont tous deux des algorithmes à source unique sur le chemin le plus court. Lorsque vous testez avec votre propre ensemble de données, essayez les deux algorithmes et comparez les résultats. Le delta-step est souvent plus rapide que celui de Bellman-Ford en raison de sa faible complexité de calcul. Toutefois, les performances dépendent du jeu de données et des paramètres d'entrée, en particulier du `delta` paramètre.

Optimisez les écritures

Suivez les pratiques suivantes pour optimiser les opérations d'écriture dans Neptune Analytics :

- Recherchez le moyen le plus efficace de charger des données dans un graphique. Lorsque vous chargez à partir de données dans Amazon S3, utilisez l'importation en bloc si la taille des données est supérieure à 50 Go. Pour des données plus petites, utilisez le chargement par lots. Si vous rencontrez out-of-memory des erreurs lorsque vous exécutez le chargement par lots, pensez à augmenter la valeur m-NCU ou à diviser la charge en plusieurs demandes. L'un des moyens d'y parvenir consiste à répartir les fichiers entre plusieurs préfixes dans le compartiment S3. Dans ce cas, appelez batch load séparément pour chaque préfixe.
- Utilisez l'importation en bloc ou le chargeur par lots pour renseigner l'ensemble initial de données graphiques. Utilisez les opérations transactionnelles de création, de mise à jour et de suppression d'OpenCypher uniquement pour les modifications mineures.

- Utilisez l'importation en bloc ou le chargeur par lots avec une simultanéité de 1 (thread unique) pour intégrer les éléments intégrés dans le graphique. Essayez de charger les intégrations dès le départ en utilisant l'une de ces méthodes.
- Évaluez la dimension des intégrations vectorielles nécessaires pour une recherche de similarité précise dans les algorithmes de recherche de similarité vectorielle. Utilisez une dimension plus petite si possible. Cela se traduit par une vitesse de chargement plus rapide pour les intégrations.
- Utilisez des algorithmes de mutation pour mémoriser les résultats algorithmiques si nécessaire. Par exemple, l'algorithme de centralité Degré Mutate trouve le degré de chaque nœud d'entrée et écrit cette valeur en tant que propriété du nœud. Si les connexions entourant ces nœuds ne changent pas par la suite, la propriété contient le résultat correct. Il n'est pas nécessaire de réexécuter l'algorithme.
- Utilisez l'action administrative de réinitialisation du graphe pour effacer tous les nœuds, arêtes et intégrations si vous devez recommencer à zéro. Il n'est pas possible de supprimer tous les nœuds, arêtes et intégrations à l'aide d'une requête OpenCypher si votre graphe est volumineux. Le délai d'expiration d'une seule requête sur un ensemble de données volumineux peut être dépassé. À mesure que la taille augmente, le jeu de données prend plus de temps à être supprimé et la taille des transactions augmente. En revanche, le temps nécessaire pour effectuer une réinitialisation du graphe est à peu près constant, et l'action permet de créer un instantané avant de l'exécuter.

Graphiques à la bonne taille

Les performances globales dépendent de la capacité allouée à un graphe Neptune Analytics. La capacité est mesurée en unités appelées unités de capacité Neptune optimisées pour la mémoire (m -). NCUs Assurez-vous que la taille de votre graphique est suffisante pour prendre en charge la taille de votre graphique et vos requêtes. Notez que l'augmentation de la capacité n'améliore pas nécessairement les performances d'une requête individuelle.

Si possible, créez le graphique en important des données depuis une source existante telle qu'Amazon S3 ou un cluster ou un instantané Neptune existant. Vous pouvez limiter la capacité minimale et maximale. Vous pouvez également modifier la capacité allouée sur un graphique existant.

Surveillez CloudWatch des indicateurs tels que `NumQueuedRequestsPerSec`,`NumOpenCypherRequestsPerSec`,`GraphStorageUsagePercent`,`Graph` et `CPUUtilization` pour déterminer si le graphique est de bonne taille. Déterminez si une capacité supplémentaire est nécessaire pour supporter la taille et la charge de votre graphe. Pour plus

d'informations sur la manière d'interpréter certains de ces indicateurs, consultez la section Pilier d'excellence opérationnelle.

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 Neptune Analytics.

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

Avant d'adopter Neptune Analytics, déterminez si votre cas d'utilisation convient à l'analyse graphique.

- Bases de données de graphes : une base de données de graphes telle que 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. Dans ce cas d'utilisation, déterminez s'il serait préférable de représenter vos données sous forme d'attributs d'un objet.
  - Données JSON ou objets binaires de grande taille (blob) stockées sous forme de propriétés.

- Analyse graphique : Neptune Analytics est un moteur de base de données d'analyse graphique capable d'analyser rapidement de grandes quantités de données graphiques en mémoire pour obtenir des informations et identifier des tendances. Vous pouvez stocker et interroger des données graphiques à la fois dans une base de données Neptune et dans un graphe Neptune Analytics. Une base de données Neptune est la mieux adaptée aux besoins de traitement transactionnel en ligne (OLTP) évolutif. Neptune Analytics est la solution idéale pour les charges de travail analytiques éphémères. Vous pouvez utiliser les deux en combinaison en chargeant les données de votre base de données Neptune axée sur les transactions vers un graphe Neptune Analytics afin d'analyser ces données. Lorsque l'analyse est terminée, vous pouvez supprimer le graphe Neptune Analytics. Pour une comparaison plus détaillée, voir Quand utiliser Neptune

Analytics et quand utiliser la base de données Neptune dans la documentation de Neptune Analytics.

Déterminez, en tenant compte des coûts, la meilleure façon de remplir votre graphique Neptune Analytics.

- Importez en bloc des données graphiques stockées dans un compartiment S3. Nous recommandons cette option si vos données ont déjà été préparées pour être chargées en masse dans une base de données Neptune, ou si vous possédez déjà, ou si vous pouvez facilement produire, les données à analyser au format CSV ou dans d'autres formats pris en charge requis par l'importation en masse. Vous pouvez exécuter l'importation en bloc dans le cadre de la procédure de création du graphe. Vous pouvez limiter la capacité minimale et maximale. Vous pouvez également exécuter l'importation sur un graphique vide créé précédemment et surveiller la tâche d'importation pendant son exécution.
- Vous pouvez créer un graphique vide, puis le remplir via une requête OpenCypher en utilisant le chargement par lots. Cette option est idéale si les données à charger sont stockées dans Amazon S3 et que leur taille est inférieure à 50 Go.
- Vous pouvez remplir le graphique à partir des données de votre cluster de bases de données Neptune (pris en charge dans la version 1.3.0 ou ultérieure de la base de données Neptune). Le but de ce modèle est d'exécuter des analyses sur les données qui se trouvent actuellement dans votre base de données de graphes. Même si la base de données a été initialement remplie par chargement groupé, elle a peut-être changé de manière significative depuis lors. Pour effectuer une importation depuis la base de données, Neptune Analytics clone votre base de données et exporte les données du clone vers un compartiment S3. Cette procédure entraîne des coûts : notamment les coûts liés à la base de données Neptune pour l'exécution du clone et les coûts liés au stockage et à la consommation des données exportées par Amazon S3. Le clone est supprimé lorsque l'exportation est terminée. Vous pouvez supprimer les données exportées dans Amazon S3.
- Vous pouvez remplir le graphique à partir de l'instantané d'un cluster de bases de données Neptune. Cette option est similaire à l'option précédente, sauf que la source est un instantané de base de données. Pour effectuer une importation à partir d'un instantané, Neptune Analytics restaure d'abord l'instantané dans un nouveau cluster de bases de données, puis exporte les données vers un compartiment S3. Cette procédure entraîne des coûts : notamment les coûts liés à la base de données Neptune pour l'exécution du cluster restauré et les coûts liés au stockage et à la consommation des données exportées par Amazon S3.

- Vous pouvez également exécuter des requêtes OpenCypher pour créer, mettre à jour ou supprimer des données en utilisant des transactions conformes aux normes ACID (atomicité, cohérence, isolation, durabilité) sur le graphique. Nous recommandons cette approche pour effectuer de petites mises à jour, mais pas pour ensemencer le graphique.

Si les données nécessaires à l'analyse sont déjà stockées dans Amazon S3, nous vous recommandons de les importer en bloc ou de les charger par lots. Elles sont plus économiques que le remplissage du graphique à partir d'un cluster de base de données Neptune ou d'un instantané.

Sélectionnez les ressources en tenant compte des coûts

La tarification de Neptune Analytics utilise une unité connue sous le nom d'unité de capacité Neptune optimisée pour la mémoire (m-NCU). Il existe un coût horaire fixe pour exécuter un graphique avec un m-NCU donné. Un graphe peut contenir des répliques pour le basculement, et ces répliques entraînent également un coût horaire de m-NCU.

Nous recommandons les meilleures pratiques suivantes pour estimer la capacité, limiter les coûts et surveiller les coûts par rapport aux performances :

- Si possible, créez le graphique en important des données provenant d'une source existante : des données stockées dans Amazon S3 ou un cluster ou un instantané Neptune existant. Cela vous permet d'économiser des efforts, car Neptune Analytics se charge de l'ensemencement du graphique et vous pouvez définir une capacité maximale limitée.
- Vous pouvez modifier la capacité allouée sur un graphique existant.
- Lorsque le graphique n'est plus nécessaire, vous pouvez créer un instantané et le supprimer. Si vous devez le réutiliser, vous pouvez restaurer le graphique à partir de l'instantané.
- Vous pouvez choisir le nombre de répliques lorsque vous créez le graphique. Définissez la valeur en fonction de vos exigences en matière de disponibilité des outils d'analyse. Réduisez les coûts en minimisant ce paramètre. La valeur maximale de 2 autorise deux instances de réplication dans des zones de disponibilité distinctes. La valeur minimale de 0 signifie que Neptune Analytics n'exécutera pas de réplique. Toutefois, la restauration est plus rapide lorsqu'une réplique est disponible. Pour une explication de la défaillance et de la restauration du graphe, consultez la section Pilier de fiabilité.
- Surveillez les dépenses de Neptune Analytics pour les périodes de facturation actuelles et passées en utilisant. AWS Billing and Cost Management

- Surveillez les métriques de Neptune Analytics pour CloudWatch, en particulierNumQueuedRequestsPerSec,NumOpenCypherRequestsPerSec,GraphStorageUsagePer et GraphSizeBytesCPUUtlization, afin de déterminer si la capacité allouée est correctement dimensionnée pour le graphique. Déterminez si une capacité inférieure peut répondre au taux de demandes observé, à l'utilisation du processeur et à la taille du graphique.
- Si vous avez besoin d'un point de terminaison privé pour votre graphe, faites attention aux coûts liés aux points de IPs terminaison flexibles de cloud privé virtuel (VPC), aux passerelles NAT ou aux autres coûts liés au VPC. Pour en savoir plus, consultez les tarifs Amazon VPC et Amazon EC2.
- Vous souhaiterez peut-être exécuter une ou plusieurs instances de Neptune Notebook afin de fournir une interface client permettant aux développeurs et aux analystes d'interroger et de visualiser le graphique (voir la tarification de Neptune Workbench). Pour minimiser les coûts, partagez l'instance entre les utilisateurs et créez des dossiers de bloc-notes distincts pour chaque utilisateur. Arrêtez l'instance lorsqu'elle n'est pas utilisée. Pour une approche permettant d'automatiser l'arrêt, consultez le billet de AWS blog Automatiser l'arrêt et le démarrage des ressources de l'environnement Amazon Neptune à l'aide de balises de ressources.

Pilier de durabilité

Le pilier du développement durable du AWS Well-Architected Framework 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 logicielles plus efficaces
- Utilisation de services gérés
- Réduction de l'impact en aval

Ce guide vise à comprendre 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 qui auront un impact environnemental positif si elles sont adoptées dans la conception de votre charge de travail et dans les opérations en cours.

Tenez compte de votre 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 Analytics est disponible.

Optimisez la consommation

Réduisez la consommation de Neptune Analytics en pratiquant les méthodes suivantes :

- Les analyses sont souvent éphémères. Le graphique n'est nécessaire que pour le temps nécessaire à l'exécution des algorithmes et à l'enregistrement des résultats. Si tel est le cas, capturez le graphique et supprimez-le lorsqu'il n'est plus nécessaire. Vous pouvez le restaurer ultérieurement à partir d'un instantané si nécessaire.
- Si la charge de travail est éphémère et que vous avez la possibilité de décider à quel moment exécuter les analyses, tenez compte des day-to-day tendances en matière de consommation d'énergie. La demande d'électricité est plus élevée à certaines périodes. Si vous êtes aux États-Unis, consultez les indicateurs de consommation quotidienne d'électricité sur le site Web de l'Energy Information Administration (EIA) des États-Unis. Exécutez des charges de travail pendant les périodes creuses dans votre région, si possible.
- Si la charge de travail n'est pas éphémère mais ne doit être disponible que pendant des périodes limitées, supprimez le graphique et restaurez-le à partir d'un instantané lorsque cela est nécessaire. Si sa disponibilité suit un calendrier, automatisez le processus de restauration à l'aide de scripts afin que le graphique soit prêt à l'heure prévue.
- Si les données sont en lecture seule ou n'ont pas changé depuis le dernier instantané, ne les capturez pas à nouveau avant de les supprimer.
- Arrêtez les blocs-notes Neptune lorsqu'ils ne sont pas utilisés.
- Surveillez CloudWatch des indicateurs tels que `NumQueuedRequestsPerSec`,`NumOpenCypherRequestsPerSec`,`GraphStorageUsagePercent`,`Grap` et `CPUUtilization` pour déterminer si le graphique est surdimensionné. Déterminez si une capacité d'instance inférieure peut s'adapter au taux de demandes observé, à l'utilisation du processeur et à la taille du graphique.

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 :

- Optimisez les requêtes et les invocations d'algorithmes graphiques. Utilisez des requêtes paramétrées et utilisez le cache du plan de requêtes, qui est activé par défaut. Pour les requêtes

lentes, exécutez un plan d'explication pour apporter des améliorations. Si vous utilisez la recherche par similarité vectorielle, déterminez si les petites incorporations produisent des résultats de similarité précis, car les petites incorporations peuvent être créées, stockées et recherchées plus efficacement. Avant d'appeler un algorithme graphique, utilisez une `MATCH` clause pour minimiser l'ensemble de nœuds en entrée. Filtrez sur les étiquettes des nœuds et des arêtes si possible.

- Recherchez le moyen le plus efficace de charger des données dans le graphique. Si vous chargez à partir de données dans Amazon S3, utilisez l'importation en bloc si la taille des données est supérieure à 50 Go. Utilisez le chargement par lots pour des données plus petites.
- Demandez aux développeurs de partager les instances du bloc-notes Neptune au lieu de créer chacun sa propre instance. Créez des dossiers de bloc-notes distincts pour chaque développeur sur une seule instance Jupyter. Arrêtez l'instance lorsqu'elle n'est pas utilisée.

Ressources

Références

- AWS Well-Architected
- AWS Documentation du framework Well-Architected
- Appliquer le framework AWS Well-Architected pour Amazon Neptune
- Meilleures pratiques de Neptune Analytics

Articles de blog et vidéos

- Articles du blog Neptune (blog de AWS base de données)
- Automatisez l'arrêt et le démarrage des ressources de l'environnement Amazon Neptune à l'aide de balises de ressources (blog de AWS base de données)
- Snackables Amazon Neptune (courtes vidéos) YouTube

Entraînement

- Commencer à utiliser Amazon Neptune
- Créez avec Amazon Neptune
- Modélisation des données pour Amazon Neptune
- Atelier Amazon Neptune Analytics

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
Publication initiale	—	20 décembre 2024

AWS Glossaire des directives prescriptives

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

Nombres

7 R

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

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

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

A

ABAC

Voir contrôle d'accès basé sur les attributs.

services abstraits

Consultez la section Services gérés.

ACIDE

Voir atomicité, consistance, isolation, durabilité.

migration active-active

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

migration active-passive

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

fonction d'agrégation

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

AI

Voir intelligence artificielle.

AIOps

Voir les opérations d'intelligence artificielle.

anonymisation

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

anti-motif

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

contrôle des applications

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

portefeuille d'applications

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

intelligence artificielle (IA)

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

opérations d'intelligence artificielle (AIOps)

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

chiffrement asymétrique

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

atomicité, cohérence, isolement, durabilité (ACID)

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

contrôle d'accès par attributs (ABAC)

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

source de données faisant autorité

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

Zone de disponibilité

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

AWS Cadre d'adoption du cloud (AWS CAF)

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

l'organisation à une adoption réussie du cloud. Pour plus d'informations, veuillez consulter le site Web AWS CAF et le livre blanc AWS CAF.

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

Outil qui évalue les charges de travail liées à la migration des bases de données, recommande des stratégies de migration et fournit des estimations de travail. AWS Le WQF est inclus avec AWS Schema Conversion Tool (AWS SCT). Il analyse les schémas de base de données et les objets de code, le code d'application, les dépendances et les caractéristiques de performance, et fournit des rapports d'évaluation.

B

mauvais bot

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

BCP

Consultez la section Planification de la continuité des activités.

graphique de comportement

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

système de poids fort

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

classification binaire

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

filtre de Bloom

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

déploiement bleu/vert

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

bot

Application logicielle qui exécute des tâches automatisées sur Internet et simule l'activité ou l'interaction humaine. Certains robots sont utiles ou bénéfiques, comme les robots d'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

Voir base de données de gestion de configuration.

référentiel de code

Emplacement où le code source et d'autres ressources, comme la documentation, les exemples et les scripts, sont stockés et mis à jour par le biais de processus de contrôle de version. Les référentiels cloud courants incluent GitHub 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 is commonly described as a pipeline. CI/CDpeut vous aider à automatiser les processus, à améliorer la productivité, à améliorer la qualité du code et à accélérer les livraisons. Pour plus d'informations, veuillez consulter Avantages de la livraison continue. CD peut également signifier déploiement continu. Pour plus d'informations, veuillez consulter Livraison continue et déploiement continu.

CV

Voir vision par ordinateur.

D

données au repos

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

classification des données

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

dérive des données

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

données en transit

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

maillage de données

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

minimisation des données

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

périmètre de données

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

prétraitement des données

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

provenance des données

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

sujet des données

Personne dont les données sont collectées et traitées.

entrepôt des données

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

langage de définition de base de données (DDL)

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

langage de manipulation de base de données (DML)

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

DDL

Voir langage de définition de base de données.

ensemble profond

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

deep learning

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

defense-in-depth

Approche de la sécurité de l'information dans laquelle une série de mécanismes et de contrôles de sécurité sont judicieusement répartis sur l'ensemble d'un réseau informatique afin de protéger la confidentialité, l'intégrité et la disponibilité du réseau et des données qu'il contient. Lorsque vous adoptez cette stratégie AWS, vous ajoutez plusieurs contrôles à différentes couches de

la AWS Organizations structure afin de sécuriser les ressources. Par exemple, une defense-in-depth approche peut combiner l'authentification multifactorielle, la segmentation du réseau et le chiffrement.

administrateur délégué

Dans AWS Organizations, un service compatible peut enregistrer un compte AWS membre pour administrer les comptes de l'organisation et gérer les autorisations pour ce service. Ce compte est appelé administrateur délégué pour ce service. Pour plus d'informations et une liste des services compatibles, veuillez consulter la rubrique Services qui fonctionnent avec AWS Organizations dans la documentation AWS Organizations .

déploiement

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

environnement de développement

Voir environnement.

contrôle de détection

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

cartographie de la chaîne de valeur du développement (DVSM)

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

jumeau numérique

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

tableau des dimensions

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

catastrophe

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

reprise après sinistre (DR)

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

DML

Voir langage de manipulation de base de données.

conception axée sur le domaine

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

DR

Voir reprise après sinistre.

détection de dérive

Suivi des écarts par rapport à une configuration de référence. Par exemple, vous pouvez l'utiliser AWS CloudFormation pour détecter la dérive des ressources du système ou AWS Control Tower

pour détecter les modifications de votre zone d'atterrissage susceptibles d'affecter le respect des exigences de gouvernance.

DVSM

Voir la cartographie de la chaîne de valeur du développement.

E

EDA

Voir analyse exploratoire des données.

EDI

Voir échange de données informatisé.

informatique de périphérie

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

échange de données informatisé (EDI)

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

chiffrement

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

clé de chiffrement

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

endianisme

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

point de terminaison

Voir point de terminaison de service.

service de point de terminaison

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

planification des ressources d'entreprise (ERP)

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

chiffrement d'enveloppe

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

environment

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

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

environnements de préproduction et des environnements pour les tests d'acceptation par les utilisateurs.

époppée

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

ERP

Voir Planification des ressources d'entreprise.

analyse exploratoire des données (EDA)

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

F

tableau des faits

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

échouer rapidement

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

limite d'isolation des défauts

Dans le AWS Cloud, une limite telle qu'une zone de disponibilité Région AWS, un plan de contrôle ou un plan de données qui limite l'effet d'une panne et contribue à améliorer la résilience des

charges de travail. Pour plus d'informations, consultez la section Limites d'isolation des AWS pannes.

branche de fonctionnalités

Voir succursale.

fonctionnalités

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

importance des fonctionnalités

Le niveau d'importance d'une fonctionnalité pour les prédictions d'un modèle. Il s'exprime généralement sous la forme d'un score numérique qui peut être calculé à l'aide de différentes techniques, telles que la méthode Shapley Additive Explanations (SHAP) et les gradients intégrés. Pour plus d'informations, voir Interprétabilité du modèle d'apprentissage automatique avec AWS.

transformation de fonctionnalité

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

invitation en quelques coups

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

FGAC

Découvrez le contrôle d'accès détaillé.

contrôle d'accès détaillé (FGAC)

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

migration instantanée (flash-cut)

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

FM

Voir le modèle de fondation.

modèle de fondation (FM)

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

G

IA générative

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

blocage géographique

Voir les restrictions géographiques.

restrictions géographiques (blocage géographique)

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

Flux de travail Gitflow

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

image dorée

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

stratégie inédite

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

barrière de protection

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

H

HA

Découvrez la haute disponibilité.

migration de base de données hétérogène

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

haute disponibilité (HA)

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

modernisation des historiens

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

données de rétention

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

migration de base de données homogène

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

données chaudes

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

correctif

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

période de soins intensifs

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

I

IaC

Considérez l'infrastructure comme un code.

politique basée sur l’identité

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

application inactive

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

IIo T

Voir Internet industriel des objets.

infrastructure immuable

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

VPC entrant (d'entrée)

Dans une architecture AWS multi-comptes, un VPC qui accepte, inspecte et achemine les connexions réseau depuis l'extérieur d'une application. L'architecture AWS de référence de sécurité recommande de configurer votre compte réseau avec les fonctions entrantes, sortantes

et d'inspection VPCs afin de protéger l'interface bidirectionnelle entre votre application et l'Internet en général.

migration incrémentielle

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

Industry 4.0

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

infrastructure

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

infrastructure en tant que code (IaC)

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

Internet industriel des objets (IIoT)

L'utilisation de capteurs et d'appareils connectés à Internet dans les secteurs industriels tels que la fabrication, l'énergie, l'automobile, les soins de santé, les sciences de la vie et l'agriculture. Pour plus d'informations, voir Élaboration d'une stratégie de transformation numérique de l'Internet des objets (IIoT) industriel.

VPC d'inspection

Dans une architecture AWS multi-comptes, un VPC centralisé qui gère les inspections du trafic réseau VPCs entre (identique ou Régions AWS différent), Internet et les réseaux locaux. L'architecture AWS de référence de sécurité recommande de configurer votre compte réseau

avec les fonctions entrantes, sortantes et d'inspection VPCs afin de protéger l'interface bidirectionnelle entre votre application et l'Internet en général.

Internet des objets (IoT)

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

interprétabilité

Caractéristique d'un modèle de machine learning qui décrit dans quelle mesure un être humain peut comprendre comment les prédictions du modèle dépendent de ses entrées. Pour plus d'informations, voir Interprétabilité du modèle d'apprentissage automatique avec AWS.

IoT

Voir Internet des objets.

Bibliothèque d'informations informatiques (ITIL)

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

gestion des services informatiques (ITSM)

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

ITIL

Consultez la bibliothèque d'informations informatiques.

ITSM

Voir Gestion des services informatiques.

L

contrôle d'accès basé sur des étiquettes (LBAC)

Une implémentation du contrôle d'accès obligatoire (MAC) dans laquelle une valeur d'étiquette de sécurité est explicitement attribuée aux utilisateurs et aux données elles-mêmes. L'intersection

entre l'étiquette de sécurité utilisateur et l'étiquette de sécurité des données détermine les lignes et les colonnes visibles par l'utilisateur.

zone de destination

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

grand modèle de langage (LLM)

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

migration de grande envergure

Migration de 300 serveurs ou plus.

LBAC

Voir contrôle d'accès basé sur des étiquettes.

principe de moindre privilège

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

lift and shift

Voir 7 Rs.

système de poids faible

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

LLM

Voir le grand modèle de langage.

environnements inférieurs

Voir environnement.

M

machine learning (ML)

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

branche principale

Voir succursale.

malware

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

services gérés

Services AWS pour lequel AWS fonctionnent la couche d'infrastructure, le système d'exploitation et les plateformes, et vous accédez aux points de terminaison pour stocker et récupérer des données. Amazon Simple Storage Service (Amazon S3) et Amazon DynamoDB sont des exemples de services gérés. Ils sont également connus sous le nom de services abstraits.

système d'exécution de la fabrication (MES)

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

MAP

Voir Migration Acceleration Program.

mécanisme

Processus complet au cours duquel vous créez un outil, favorisez son adoption, puis inspectez les résultats afin de procéder aux ajustements nécessaires. Un mécanisme est un cycle qui se renforce et s'améliore au fur et à mesure de son fonctionnement. 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 avec le service de migration AWS d'applications.

Évaluation du portefeuille de migration (MPA)

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

Évaluation de la préparation à la migration (MRA)

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

stratégie de migration

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

ML

Voir apprentissage automatique.

modernisation

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

évaluation de la préparation à la modernisation

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

applications monolithiques (monolithes)

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

MPA

Voir Évaluation du portefeuille de migration.

MQTT

Voir Message Queuing Telemetry Transport.

classification multi-classes

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

infrastructure mutable

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

O

OAC

Voir Contrôle d'accès à l'origine.

OAI

Voir l'identité d'accès à l'origine.

OCM

Voir gestion du changement organisationnel.

migration hors ligne

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

OI

Voir Intégration des opérations.

OLA

Voir l'accord au niveau opérationnel.

migration en ligne

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

OPC-UA

Voir Open Process Communications - Architecture unifiée.

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

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

accord au niveau opérationnel (OLA)

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

examen de l'état de préparation opérationnelle (ORR)

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

technologie opérationnelle (OT)

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

intégration des opérations (OI)

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

journal de suivi d'organisation

Un parcours créé par AWS CloudTrail qui enregistre tous les événements pour tous les membres Comptes AWS d'une organisation dans AWS Organizations. Ce journal de suivi est créé dans

chaque Compte AWS qui fait partie de l'organisation et suit l'activité de chaque compte. Pour plus d'informations, consultez la section Création d'un suivi pour une organisation dans la CloudTrail documentation.

gestion du changement organisationnel (OCM)

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

contrôle d'accès d'origine (OAC)

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

identité d'accès d'origine (OAI)

Dans CloudFront, une option permettant de restreindre l'accès afin de sécuriser votre contenu Amazon S3. Lorsque vous utilisez OAI, il CloudFront crée un principal auprès duquel Amazon S3 peut s'authentifier. Les principaux authentifiés ne peuvent accéder au contenu d'un compartiment S3 que par le biais d'une distribution spécifique CloudFront . Voir également OAC, qui fournit un contrôle d'accès plus précis et amélioré.

ORR

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

DE

Voir technologie opérationnelle.

VPC sortant (de sortie)

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

P

limite des autorisations

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

informations personnelles identifiables (PII)

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

PII

Voir les informations personnelles identifiables.

manuel stratégique

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

PLC

Voir contrôleur logique programmable.

PLM

Consultez la section Gestion du cycle de vie des produits.

politique

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

persistance polyglotte

Choix indépendant de la technologie de stockage de données d'un microservice en fonction des modèles d'accès aux données et d'autres exigences. Si vos microservices utilisent la même

technologie de stockage de données, ils peuvent rencontrer des difficultés d'implémentation ou présenter des performances médiocres. Les microservices sont plus faciles à mettre en œuvre, atteignent de meilleures performances, ainsi qu'une meilleure capacité de mise à l'échelle s'ils utilisent l'entrepôt de données le mieux adapté à leurs besoins. Pour plus d'informations, veuillez consulter Enabling data persistence in microservices.

évaluation du portefeuille

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

predicate

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

prédicat pushdown

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

contrôle préventif

Contrôle de sécurité conçu pour empêcher qu'un événement ne se produise. Ces contrôles constituent une première ligne de défense pour empêcher tout accès non autorisé ou toute modification indésirable de votre réseau. Pour plus d'informations, veuillez consulter Preventative controls dans Implementing security controls on AWS.

principal

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

confidentialité dès la conception

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

zones hébergées privées

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

VPCs domaines. Pour plus d'informations, veuillez consulter Working with private hosted zones dans la documentation Route 53.

contrôle proactif

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

gestion du cycle de vie des produits (PLM)

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

environnement de production

Voir environnement.

contrôleur logique programmable (PLC)

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

chaînage rapide

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

pseudonymisation

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

publish/subscribe (pub/sub)

Modèle qui permet des communications asynchrones entre les microservices afin d'améliorer l'évolutivité et la réactivité. Par exemple, dans un MES basé sur des microservices, un microservice peut publier des messages d'événements sur un canal auquel d'autres microservices

peuvent s'abonner. Le système peut ajouter de nouveaux microservices sans modifier le service de publication.

Q

plan de requête

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

régression du plan de requêtes

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

R

Matrice RACI

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

CHIFFON

Voir Retrieval Augmented Generation.

rançongiciel

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

Matrice RASCI

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

RCAC

Voir contrôle d'accès aux lignes et aux colonnes.

réplica en lecture

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

réarchitecte

Voir 7 Rs.

objectif de point de récupération (RPO)

Durée maximale acceptable depuis le dernier point de récupération des données. Il détermine ce qui est considéré comme étant une perte de données acceptable entre le dernier point de reprise et l'interruption du service.

objectif de temps de récupération (RTO)

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

refactoriser

Voir 7 Rs.

Région

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

régression

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

réhéberger

Voir 7 Rs.

version

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

déplacer

Voir 7 Rs.

replateforme

Voir 7 Rs.

rachat

Voir 7 Rs.

résilience

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

runbook

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

S

SAML 2.0

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

SCADA

Voir Contrôle de supervision et acquisition de données.

SCP

Voir la politique de contrôle des services.

secret

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

sécurité dès la conception

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

contrôle de sécurité

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

renforcement de la sécurité

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

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

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

automatisation des réponses de sécurité

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

chiffrement côté serveur

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

Politique de contrôle des services (SCP)

Politique qui fournit un contrôle centralisé des autorisations pour tous les comptes d'une organisation dans AWS Organizations. SCPs définissez des garde-fous ou des limites aux actions qu'un administrateur peut déléguer à des utilisateurs ou à des rôles. Vous pouvez les utiliser SCPs comme listes d'autorisation ou de refus pour spécifier les services ou les actions autorisés ou interdits. Pour plus d'informations, consultez la section Politiques de contrôle des services dans la AWS Organizations documentation.

point de terminaison du service

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

contrat de niveau de service (SLA)

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

indicateur de niveau de service (SLI)

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

objectif de niveau de service (SLO)

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

modèle de responsabilité partagée

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

SIEM

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

point de défaillance unique (SPOF)

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

SLA

Voir le contrat de niveau de service.

SLI

Voir l'indicateur de niveau de service.

SLO

Voir l'objectif de niveau de service.

split-and-seed modèle

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

SPOF

Voir point de défaillance unique.

schéma en étoile

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

modèle de figuier étrangleur

Approche de modernisation des systèmes monolithiques en réécrivant et en remplaçant progressivement les fonctionnalités du système jusqu'à ce que le système hérité puisse être mis hors service. Ce modèle utilise l'analogie d'un figuier de vigne qui se développe dans un arbre existant et qui finit par supplanter son hôte. Le schéma a été présenté par Martin Fowler comme un moyen de gérer les risques lors de la réécriture de systèmes monolithiques. Pour obtenir un

exemple d'application de ce modèle, veuillez consulter Modernizing legacy Microsoft ASP.NET (ASMX) web services incrementally by using containers and Amazon API Gateway.

sous-réseau

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

contrôle de supervision et acquisition de données (SCADA)

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

chiffrement symétrique

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

tests synthétiques

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

invite du système

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

T

balises

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

variable cible

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

liste de tâches

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

environnement de test

Voir environnement.

entraînement

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

passerelle de transit

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

flux de travail basé sur jonction

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

accès sécurisé

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

réglage

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

équipe de deux pizzas

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

U

incertitude

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

tâches indifférenciées

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

environnements supérieurs

Voir environnement.

V

mise à vide

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

contrôle de version

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

Appairage de VPC

Une connexion entre deux VPCs qui vous permet d'acheminer le trafic en utilisant des adresses IP privées. Pour plus d'informations, veuillez consulter la rubrique Qu'est-ce que l'appairage de VPC ? dans la documentation Amazon VPC.

vulnérabilités

Défaut logiciel ou matériel qui compromet la sécurité du système.

W

cache actif

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

données chaudes

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

fonction de fenêtre

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

charge de travail

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

flux de travail

Groupes fonctionnels d'un projet de migration chargés d'un ensemble de tâches spécifique. Chaque flux de travail est indépendant, mais prend en charge les autres flux de travail du projet.

Par exemple, le flux de travail du portefeuille est chargé de prioriser les applications, de planifier les vagues et de collecter les métadonnées de migration. Le flux de travail du portefeuille fournit ces actifs au flux de travail de migration, qui migre ensuite les serveurs et les applications.

VER

Voir écrire une fois, lire plusieurs.

WQF

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

écrire une fois, lire plusieurs (WORM)

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

Z

exploit Zero-Day

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

vulnérabilité « jour zéro »

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

invite Zero-Shot

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

application zombie

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

Les traductions sont fournies par des outils de traduction automatique. En cas de conflit entre le contenu d'une traduction et celui de la version originale en anglais, la version anglaise prévaudra.