Mise à l'échelle de l'infrastructure Amazon EKS pour optimiser le calcul, les charges de travail et les performances du réseau

AWS Conseils prescriptifs

AWS Conseils prescriptifs: Mise à l'échelle de l'infrastructure Amazon EKS pour optimiser le calcul, les charges de travail et les performances du réseau

Table of Contents

Mise à l'échelle de l'infrastructure Amazon EKS pour optimiser le calcul, les charges de travail et les performances du réseau

Aniket Dekate, Aniket Kurzadkar et Ishwar Chauthaiwale, Amazon Web Services ()AWS

Novembre 2024 (historique du document)

Amazon Elastic Kubernetes Service (Amazon EKS) est un service Kubernetes géré. Avec Amazon EKS, vous pouvez exécuter des pods Kubernetes dans un environnement cloud conteneurisé sans avoir à installer et à exploiter votre propre plan de contrôle. Grâce à la AWS gestion du plan de contrôle, Amazon EKS réduit la gestion opérationnelle organisationnelle. Les autres avantages de l'utilisation d'Amazon EKS incluent l'évolutivité, la fiabilité et la sécurité dans l'environnement cloud.

Ce guide est conçu pour aider les entreprises à optimiser leur infrastructure Amazon EKS dans les domaines suivants :

- La mise à l'échelle du calcul est un élément essentiel des performances des applications dans un environnement Kubernetes dynamique :
  - Allocation efficace des ressources : découvrez les techniques d'allocation dynamique des ressources calculées pour répondre à une demande variable.
  - Outils d'automatisation : obtenez une vue d'ensemble des outils et services qui automatisent la mise à l'échelle du calcul, réduisant ainsi le besoin d'intervention manuelle.
- Le dimensionnement de la charge de travail permet de s'assurer que les applications peuvent gérer différentes charges de travail sans dégrader les performances :
  - Autoscaler à modules horizontaux — Découvrez en détail comment un HPA aide à dimensionner les charges de travail en fonction de métriques en temps réel.
  - Autoscaleur proportionnel au cluster : découvrez comment le CPA adapte automatiquement et maintient une relation proportionnelle entre les nœuds et les répliques, en augmentant ou en diminuant les charges de travail en fonction de l'évolution de la taille du cluster.
  - Dimensionnement piloté par les événements : passez en revue les stratégies de dimensionnement des applications en réponse à des événements ou à des déclencheurs spécifiques.

- La mise à l'échelle du réseau permet de maintenir une communication fluide entre les services et un flux de données efficace dans des environnements dynamiques :
  - Plug-in Amazon VPC CNI : découvrez comment le plug-in VPC CNI permet une mise en réseau évolutive au sein de clusters Amazon EKS.
  - Mise en réseau personnalisée : passez en revue la gestion des adresses IP et la ségrégation du trafic réseau sur les clusters Amazon EKS.
  - Délégation de préfixes : découvrez comment rationaliser la gestion des adresses IP dans les clusters Amazon EKS de grande taille et évolutifs.
  - Amazon VPC Lattice — Découvrez comment VPC Lattice peut gérer le cross-VPC et le réseau pour une mise à l'échelle fluide. service-to-service
- L'optimisation des coûts aide les entreprises à voir où leurs ressources sont dépensées et à affecter les dépenses de manière appropriée aux départements ou aux projets :
  - Dimensionnement correct des ressources : envisagez des techniques permettant de dimensionner les ressources du cloud de manière appropriée à la charge de travail.
  - Surveillance et contrôle des coûts : passez en revue les outils et les meilleures pratiques pour suivre et optimiser les dépenses liées au cloud.

Chaque section met l'accent sur les objectifs spécifiques nécessaires pour créer un environnement cloud fiable, efficace et abordable.

Objectifs

Ce guide peut vous aider, vous et votre organisation, à atteindre les objectifs commerciaux suivants :

- Efficacité accrue des ressources : optimisez l'utilisation des ressources en adaptant dynamiquement le calcul, les charges de travail et les ressources réseau en fonction des demandes en temps réel.

  Cet objectif souligne l'importance d'augmenter ou de réduire les ressources en fonction des modèles d'utilisation réels. Des outils tels que les autoscalers à modules horizontaux et le plug-in Amazon VPC CNI aident les entreprises à n'utiliser que les ressources dont elles ont besoin, en minimisant le gaspillage et en optimisant les performances.

- Performances améliorées des applications — Maintenez des performances et une réactivité élevées des applications, même en cas de charges de travail et de modèles de trafic fluctuants.

Cet objectif met l'accent sur les stratégies visant à garantir que les applications peuvent gérer les pics de trafic et les charges de travail élevées sans compromettre les performances. Des techniques telles que le dimensionnement de la charge de travail piloté par les événements, l'allocation de calcul efficace et les architectures réseau évolutives sont essentielles pour atteindre cet objectif.

- Évolutivité sans faille — Facilitez la mise à l'échelle des composants de l'infrastructure, ce qui permet une croissance et une adaptation sans effort aux besoins changeants de l'entreprise.

  Une évolutivité sans faille est essentielle pour les entreprises qui anticipent une croissance ou sont confrontées à des niveaux de trafic variables. Cet objectif tient compte de l'importance de mettre en œuvre des solutions évolutives pour les ressources de calcul, de charge de travail et de réseau, afin que le dimensionnement puisse être automatique, efficace et transparent.

- Optimisation des coûts : minimisez les coûts du cloud tout en maintenant ou en améliorant les performances et l'évolutivité.

  L'optimisation des coûts peut inclure la réduction des dépenses, par exemple en adaptant les ressources, en utilisant des solutions de mise à l'échelle rentables et en surveillant les dépenses. L'objectif est de trouver un équilibre entre les économies de coûts et le besoin de performances et d'évolutivité élevées.

Dimensionnement du calcul

La mise à l'échelle du calcul est un élément essentiel des performances des applications dans un environnement Kubernetes dynamique. Kubernetes réduit le gaspillage grâce à l'ajustement dynamique des ressources informatiques (telles que le processeur et la mémoire) en réponse à la demande en temps réel. Cette fonctionnalité permet d'éviter le surprovisionnement ou le sous-provisionnement, ce qui peut également réduire les dépenses d'exploitation. Kubernetes élimine efficacement le besoin d'intervention manuelle en permettant à l'infrastructure de s'étendre automatiquement pendant les heures de pointe et de diminuer pendant les périodes creuses.

La mise à l'échelle globale du calcul de Kubernetes automatise le processus de dimensionnement, ce qui accroît la flexibilité et l'évolutivité de l'application et améliore son comportement tolérant aux pannes. En fin de compte, les capacités de Kubernetes améliorent l'excellence opérationnelle et la productivité.

Cette section décrit les types de dimensionnement de calcul suivants :

- Autoscaleur de clusters
- Cluster Autoscaler avec surprovisionnement
- Charpentier

Cluster AutoScaler

En fonction des besoins des pods, l'outil Cluster Autoscaler modifie automatiquement la taille en ajoutant des nœuds lorsque cela est nécessaire ou en supprimant des nœuds lorsqu'ils ne sont pas nécessaires et sont sous-utilisés.

Considérez l'outil Cluster Autoscaler comme une solution de dimensionnement pour les charges de travail où la demande augmente progressivement et où la latence lors de la mise à l'échelle n'est pas un problème majeur.

L'outil Cluster Autoscaler fournit les fonctionnalités clés suivantes :

- Mise à l'échelle : fait évoluer les nœuds de manière dynamique vers le haut ou vers le bas en réponse aux demandes de ressources réelles.
- Planification des modules : permet de s'assurer que chaque module fonctionne et dispose des ressources dont il a besoin pour fonctionner, évitant ainsi la pénurie de ressources.

- Rentabilité : élimine les dépenses inutiles liées à l'exploitation de nœuds sous-utilisés en les éliminant.

Cluster Autoscaler avec surprovisionnement

Cluster Autoscaler doté de fonctions de surprovisionnement similaires au Cluster Autoscaler, en ce sens qu'il déploie les nœuds de manière efficace et permet de gagner du temps en exécutant des pods de faible priorité sur les nœuds. Grâce à cette technique, le trafic est redirigé vers ces modules en réponse à des pics soudains de demande, ce qui permet à l'application de continuer à fonctionner sans interruption.

Cluster Autoscaler avec surprovisionnement offre les fonctionnalités de modules factices qui peuvent être utilisés pour déployer et exécuter facilement des nœuds lorsque la charge de travail est très importante, que la latence n'est pas nécessaire et que le dimensionnement doit être rapide.

Cluster Autoscaler avec surprovisionnement fournit les fonctionnalités clés suivantes :

- Meilleure réactivité : en rendant la capacité excédentaire accessible en permanence, le développement du cluster en réponse aux pics de demande prend moins de temps.
- Réservation des ressources : la gestion des pics de trafic inattendus contribue efficacement à une gestion correcte avec un minimum de temps d'arrêt.
- Mise à l'échelle fluide — La réduction des délais d'allocation des ressources facilite un processus de mise à l'échelle plus fluide.

Karpenter

Karpenter for Kubernetes surpasse l'outil Cluster Autoscaler traditionnel en termes d'open source, de performances et de personnalisation. Avec Karpenter, vous pouvez lancer automatiquement uniquement les ressources de calcul nécessaires pour répondre aux demandes de votre cluster en temps réel. Karpenter est conçu pour offrir une mise à l'échelle plus efficace et plus réactive.

Les applications dont les charges de travail sont extrêmement variables ou complexes, pour lesquelles des décisions rapides de mise à l'échelle sont essentielles, bénéficient grandement de l'utilisation de Karpenter. Il s'intègre AWS, offrant un déploiement amélioré et une optimisation de la sélection des nœuds.

Karpenter inclut les principales fonctionnalités suivantes :

- Provisionnement dynamique : Karpenter fournit les instances et les tailles adaptées à l'objectif et provisionne les nouveaux nœuds de manière dynamique en fonction des exigences particulières des pods.
- Planification avancée — À l'aide d'un placement intelligent des modules, Karpenter organise les nœuds de manière à ce que les ressources telles que le GPU, le processeur, la mémoire et le stockage soient utilisées le plus efficacement possible.
- Mise à l'échelle rapide — Karpenter peut effectuer une mise à l'échelle rapidement et réagit fréquemment en quelques secondes. Cette réactivité est utile pour les modèles de trafic soudain ou lorsque la charge de travail exige une mise à l'échelle immédiate
- Rentabilité : en choisissant avec soin l'instance la plus efficace, vous pouvez réduire les coûts d'exploitation et tirer parti des alternatives économiques supplémentaires proposées AWS, telles que les instances à la demande, les instances ponctuelles et les instances réservées.

Dimensionnement des charges

La mise à l'échelle de la charge de travail dans Kubernetes est essentielle pour maintenir les performances des applications et l'efficacité des ressources dans les environnements dynamiques. La mise à l'échelle permet de garantir que les applications peuvent gérer différentes charges de travail sans dégrader les performances. Kubernetes permet d'augmenter ou de diminuer automatiquement les ressources en fonction de mesures en temps réel, ce qui permet aux entreprises de réagir rapidement aux changements de trafic. Cette élasticité améliore non seulement l'expérience utilisateur, mais optimise également l'utilisation des ressources, contribuant ainsi à minimiser les coûts associés à des ressources sous-utilisées ou surapprovisionnées.

En outre, une mise à l'échelle efficace de la charge de travail favorise une haute disponibilité, garantissant ainsi la réactivité des applications même pendant les périodes de pointe. La mise à l'échelle de la charge de travail dans Kubernetes permet aux entreprises de mieux utiliser les ressources du cloud en ajustant dynamiquement les capacités pour répondre aux besoins actuels.

Cette section décrit les types suivants de dimensionnement de la charge de travail :

- Autoscaler à nacelle horizontale
- Autoscaleur proportionnel en cluster
- Autoscaler piloté par les événements basé sur Kubernetes

Horizontal Pod Autoscaler

L'Horizontal Pod Autoscaler (HPA) est une fonctionnalité de Kubernetes qui ajuste automatiquement le nombre de répliques de pods dans un déploiement, un contrôleur de réplication ou un ensemble dynamique, en fonction de l'utilisation observée du processeur ou d'autres mesures sélectionnées. Le HPA garantit que les applications peuvent gérer les fluctuations du trafic et des niveaux de charge de travail sans intervention manuelle. Le HPA offre un moyen de préserver des performances optimales tout en utilisant efficacement les ressources disponibles.

Dans les contextes où la demande des utilisateurs peut fluctuer considérablement au fil du temps, les applications Web, les microservices et APIs le HPA sont particulièrement utiles.

L'Autoscaler Horizontal Pod fournit les principales fonctionnalités suivantes :

- Dimensionnement automatique : HPA augmente ou diminue automatiquement le nombre de répliques de pods en réponse à des indicateurs en temps réel, garantissant ainsi que les applications peuvent évoluer pour répondre à la demande des utilisateurs.
- Décisions basées sur des métriques : par défaut, le HPA évolue en fonction de l'utilisation du processeur. Cependant, il peut également utiliser des métriques personnalisées, telles que l'utilisation de la mémoire ou des métriques spécifiques à l'application, ce qui permet d'élaborer des stratégies de dimensionnement plus personnalisées.
- Paramètres configurables : vous pouvez choisir le nombre minimum et maximum de répliques ainsi que les pourcentages d'utilisation souhaités, ce qui vous donne le pouvoir de décider de l'intensité de la mise à l'échelle.
- Intégration à Kubernetes : pour surveiller et modifier les ressources, HPA travaille en tandem avec d'autres éléments de l'écosystème Kubernetes, notamment le serveur de métriques, l'API Kubernetes et les adaptateurs de métriques personnalisés.
- Meilleure utilisation des ressources : le HPA contribue à garantir une utilisation efficace des ressources, à réduire les coûts et à améliorer les performances, en modifiant dynamiquement le nombre de pods.

Autoscaleur proportionnel en cluster

Le Cluster Proportional Autoscaler (CPA) est un composant Kubernetes conçu pour ajuster automatiquement le nombre de répliques de pods dans un cluster en fonction du nombre de nœuds disponibles. Contrairement aux autoscalers traditionnels qui évoluent en fonction des indicateurs d'utilisation des ressources (tels que le processeur et la mémoire), le CPA adapte les charges de travail proportionnellement à la taille du cluster lui-même.

Cette approche est particulièrement utile pour les applications qui doivent maintenir un certain niveau de redondance ou de disponibilité par rapport à la taille du cluster, telles que CoreDNS et d'autres services d'infrastructure. Les principaux cas d'utilisation du CPA sont les suivants :

- Surprovisionnement
- Élargir les services de plateforme de base
- Augmentez les charges de travail car le CPA ne nécessite pas de serveur de métriques ni d'adaptateur Prometheus

En automatisant le processus de dimensionnement, le CPA aide les entreprises à maintenir une répartition équilibrée de la charge de travail, à accroître l'efficacité des ressources et à s'assurer que les applications sont correctement provisionnées pour répondre à la demande des utilisateurs.

L'autoscaler proportionnel en cluster fournit les principales fonctionnalités suivantes :

- Dimensionnement basé sur les nœuds : le CPA adapte les répliques en fonction du nombre de nœuds de cluster pouvant être planifiés, ce qui permet aux applications de s'étendre ou de se contracter proportionnellement à la taille du cluster.
- Ajustement proportionnel : pour garantir que l'application peut évoluer en fonction de l'évolution de la taille du cluster, l'autoscaler établit une relation proportionnée entre le nombre de nœuds et le nombre de répliques. Cette relation est utilisée pour calculer le nombre de répliques souhaité pour une charge de travail.
- Intégration avec les composants Kubernetes : le CPA fonctionne avec des composants Kubernetes standard tels que le Horizontal Pod Autoscaler (HPA), mais se concentre spécifiquement sur le nombre de nœuds plutôt que sur les indicateurs d'utilisation des ressources. Cette intégration permet une stratégie de mise à l'échelle plus complète.
- Clients d'API Golang — Pour surveiller le nombre de nœuds et leurs cœurs disponibles, le CPA utilise des clients d'API Golang qui s'exécutent dans des pods et communiquent avec le serveur d'API Kubernetes.
- Paramètres configurables : à l'aide de a`ConfigMap`, les utilisateurs peuvent définir des seuils et des paramètres de dimensionnement que le CPA utilise pour modifier son comportement et s'assurer qu'il suit le plan de dimensionnement prévu.

Autoscaler piloté par les événements basé sur Kubernetes

L'Event Driven Autoscaler (KEDA) basé sur Kubernetes est un projet open source qui permet aux charges de travail Kubernetes d'évoluer en fonction du nombre d'événements à traiter. KEDA améliore l'évolutivité des applications en leur permettant de répondre de manière dynamique aux différentes charges de travail, en particulier celles qui sont dictées par des événements.

En automatisant le processus de dimensionnement en fonction des événements, KEDA aide les entreprises à optimiser l'utilisation des ressources, à améliorer les performances des applications et à réduire les coûts associés au surprovisionnement. Cette approche est particulièrement utile pour les applications soumises à des modèles de trafic variés, tels que les microservices, les fonctions sans serveur et les systèmes de traitement des données en temps réel.

KEDA fournit les fonctionnalités clés suivantes :

- Dimensionnement piloté par les événements — KEDA vous permet de définir des règles de dimensionnement basées sur des sources d'événements externes, telles que les files d'attente de messages, les requêtes HTTP ou les métriques personnalisées. Cette fonctionnalité permet de s'assurer que les applications évoluent en fonction de la demande en temps réel.
- Composant léger — KEDA est un composant léger à usage unique qui ne nécessite pas beaucoup de configuration ou de surcharge pour être facilement intégré dans les clusters Kubernetes existants.
- Intégration à Kubernetes — KEDA étend les capacités des composants natifs de Kubernetes, tels que le Horizontal Pod Autoscaler (HPA). KEDA ajoute des fonctionnalités de mise à l'échelle pilotées par les événements à ces composants, en les améliorant plutôt qu'en les remplaçant.
- Support de plusieurs sources d'événements — KEDA est compatible avec un large éventail de sources d'événements, y compris les plateformes de messagerie populaires telles que RabbitMQ, Apache Kafka, etc. Grâce à cette adaptabilité, vous pouvez personnaliser la mise à l'échelle en fonction de votre architecture unique axée sur les événements.
- Scalers personnalisés : à l'aide de scalers personnalisés, vous pouvez désigner des métriques spécifiques que KEDA peut utiliser pour lancer des actions de dimensionnement en réponse à une logique ou à des exigences commerciales spécifiques.
- Configuration déclarative — Conformément aux principes de Kubernetes, vous pouvez utiliser KEDA pour décrire le comportement de dimensionnement de manière déclarative en utilisant les ressources personnalisées de Kubernetes pour définir comment le dimensionnement doit se produire.

Dimensionnement du réseau

La mise à l'échelle du réseau dans Kubernetes est essentielle pour maintenir une communication fluide entre les services et garantir un flux de données efficace dans des environnements dynamiques. La mise à l'échelle de l'infrastructure réseau permet de garantir que le cluster peut gérer différents niveaux de trafic sans rencontrer de goulots d'étranglement ou de problèmes de latence. Kubernetes fournit des outils et des mécanismes permettant de dimensionner les ressources du réseau, ce qui permet aux entreprises de maintenir des performances optimales en fonction de l'évolution des modèles de trafic.

Cette élasticité de la mise à l'échelle du réseau améliore l'expérience utilisateur globale en garantissant des connexions rapides et fiables. La mise à l'échelle du réseau optimise également l'utilisation des ressources du réseau, ce qui contribue à réduire les coûts associés aux composants réseau sous-utilisés ou surchargés.

En outre, une mise à l'échelle efficace du réseau est essentielle pour garantir une disponibilité et une résilience élevées. En ajustant dynamiquement la capacité et le routage du réseau, les entreprises peuvent garantir l'accessibilité et la réactivité des services, même en période de pointe ou de pics de trafic inattendus. Cette approche permet une meilleure utilisation des ressources du réseau cloud, garantissant ainsi que l'infrastructure est toujours alignée sur les exigences actuelles.

Cette section décrit les types de mise à l'échelle du réseau suivants :

- Plug-in Amazon VPC CNI pour Kubernetes
- Réseau personnalisé
- Délégation de préfixes
- Amazon VPC Lattice

Plugin CNI Amazon VPC pour Kubernetes

Le plug-in Amazon VPC Container Network Interface (CNI) pour Kubernetes est un composant essentiel d'Amazon EKS. Le plug-in VPC CNI fournit des fonctionnalités réseau avancées en intégrant des pods Kubernetes à Amazon VPC. Avec ce plugin, une adresse IP unique est attribuée à chaque pod depuis le cloud privé virtuel (VPC), ce qui améliore l'isolation et les performances du réseau. À mesure que les clusters se développent et que les demandes du réseau fluctuent, le plug-in Amazon VPC CNI joue un rôle clé pour garantir des opérations réseau efficaces et évolutives.

Le plugin gère automatiquement l'allocation et le routage des adresses IP au sein du VPC, simplifiant ainsi la gestion du réseau et réduisant le risque de conflits IP. Il prend en charge des fonctionnalités telles que la délégation de préfixes, ce qui permet une gestion plus flexible des adresses IP.

Le plug-in VPC CNI aide les entreprises à optimiser les performances du réseau, à renforcer la sécurité et à réduire le risque d'épuisement des adresses IP. Ces fonctionnalités sont particulièrement utiles pour les environnements dynamiques à grande échelle où les demandes réseau fluctuent, tels que les architectures de microservices, les charges de travail à haute densité et les applications mutualisées.

Le plug-in Amazon VPC CNI fournit les fonctionnalités clés suivantes :

- Mise en réseau améliorée — Le plug-in VPC CNI permet à chaque pod de recevoir sa propre adresse IP directement depuis le VPC, ce qui garantit une isolation et des performances réseau solides. Cette approche est essentielle pour les charges de travail nécessitant un débit réseau élevé et une faible latence.
- Délégation de préfixes : pour surmonter les problèmes d'épuisement des adresses IP dans les grands clusters, la délégation de préfixes alloue de manière dynamique de plus grands blocs IPs aux nœuds, qui sont ensuite subdivisés pour être utilisés par les pods. Cette approche garantit une utilisation efficace des adresses IP et simplifie le dimensionnement du réseau.
- Mise en réseau personnalisée : les utilisateurs peuvent configurer des interfaces réseau personnalisées (ENIs) pour les pods, ce qui permet de répartir le trafic des pods sur plusieurs interfaces, de réduire la congestion du réseau et d'améliorer l'évolutivité.
- Support pour IPv6 : IPv6 en activant les clusters Amazon EKS, les utilisateurs peuvent étendre de manière significative l'espace d'adresses IP disponible, facilitant ainsi le dimensionnement d'applications distribuées de grande taille sans contraintes ni IPv4 limites.
- Intégration à Kubernetes — Le plug-in VPC CNI fonctionne parfaitement avec les composants réseau Kubernetes, garantissant ainsi leur gestion efficace entre les pods, les services et IPs les points de terminaison externes, et il prend en charge des fonctionnalités avancées telles que les groupes de sécurité pour les pods.

Mise en réseau personnalisée

La mise en réseau personnalisée dans Amazon EKS permet d'attribuer des interfaces réseau spécifiques aux pods, offrant ainsi un meilleur contrôle de la gestion des adresses IP et du trafic réseau. Cette approche est particulièrement utile dans les scénarios où l'épuisement des adresses IP

est préoccupant ou lorsqu'il est nécessaire de séparer le trafic réseau pour des raisons de sécurité, de conformité ou de performance. La mise en réseau personnalisée aide les entreprises à gérer efficacement l'espace d'adresses IP, à séparer le trafic et à garantir des performances réseau évolutives.

Grâce à la mise en réseau personnalisée, les administrateurs peuvent gérer les ressources du réseau plus efficacement. Les administrateurs peuvent utiliser un réseau personnalisé pour s'assurer que les pods disposent de l'isolation réseau nécessaire et que le cluster peut évoluer sans rencontrer de limites d'adresses IP.

La mise en réseau personnalisée fournit les fonctionnalités clés suivantes :

- Gestion IP améliorée — La mise en réseau personnalisée permet d'attribuer des interfaces réseau spécifiques (ENIs) aux pods, ce qui permet de gérer l'épuisement des adresses IP en répartissant le trafic des pods entre plusieurs ENIs. Cette fonctionnalité est particulièrement importante dans les clusters avec des charges de travail à haute densité.
- Ségrégation du trafic — Grâce aux interfaces réseau personnalisées, vous pouvez séparer le trafic des pods en fonction de critères spécifiques, tels que le type d'application ou les exigences de sécurité. Cette approche permet de mieux contrôler la manière dont le trafic circule à l'intérieur et à l'extérieur du cluster.
- Support pour IPv6 : la mise en réseau personnalisée d'Amazon EKS est également compatible IPv6, offrant ainsi une solution aux limites d' IPv4adresses. Le réseau peut évoluer efficacement sans conflits d'adresses IP, même dans le cadre de déploiements à grande échelle.
- Évolutivité et flexibilité — À mesure que le cluster évolue, la mise en réseau personnalisée permet une gestion dynamique des interfaces réseau. Les ressources réseau appropriées sont attribuées aux nouveaux pods sans intervention manuelle. Cette approche permet de maintenir un environnement réseau flexible et évolutif capable de s'adapter à l'évolution des charges de travail.

Délégation de préfixes

La délégation de préfixes dans Kubernetes, en particulier dans Amazon EKS, est conçue pour rationaliser et optimiser la gestion des adresses IP à mesure que les clusters évoluent. En allouant dynamiquement de plus grands blocs d'adresses IP (préfixes) aux nœuds, la délégation de préfixes réduit le risque d'épuisement des adresses IP et simplifie la gestion de l'espace IP.

Cette approche améliore l'efficacité du réseau, minimise la fragmentation et aide les clusters à évoluer en douceur sans ajustement manuel de la plage IP. La délégation de préfixes est

particulièrement utile pour les déploiements à grande échelle, les charges de travail à haute densité et les environnements dans lesquels une gestion flexible et dynamique des adresses IP est essentielle au maintien des performances et de l'évolutivité du réseau.

La délégation de préfixes fournit les fonctionnalités clés suivantes :

- Gestion efficace des adresses IP — La délégation de préfixes permet une allocation dynamique des plages d'adresses IP, réduisant ainsi le risque d'épuisement des adresses IP et garantissant une utilisation efficace de l'espace IP disponible.
- Gestion du réseau simplifiée : en permettant aux nœuds de gérer leurs propres allocations d'adresses IP, la délégation de préfixes minimise la fragmentation du réseau et simplifie le processus de routage, facilitant ainsi le dimensionnement des clusters en fonction des besoins.
- Support pour les déploiements à grande échelle — Dans les grands clusters avec des charges de travail à haute densité, la délégation de préfixes permet une mise à l'échelle fluide en permettant à de nouveaux nœuds de rejoindre le cluster sans ajustement manuel de la plage d'adresses IP.

Amazon VPC Lattice

Amazon VPC Lattice permet une service-to-service communication efficace et sécurisée au sein et entre les deux VPCs, en particulier dans les architectures de microservices. VPC Lattice utilise des mesures de sécurité telles que des groupes de sécurité et des listes de contrôle d'accès réseau (réseau ACLs) en plus de l'intégration Gestion des identités et des accès AWS (IAM) pour une authentification précise des applications. Un service proxy de couche 7 au cœur de VPC Lattice assure la connexion, l'équilibrage de charge, l'authentification, l'autorisation, l'observabilité, la gestion du trafic et la découverte de services.

En simplifiant les configurations réseau et de sécurité, VPC Lattice aide les entreprises à optimiser la gestion du trafic, à améliorer les performances des applications et à évoluer de manière fluide entre plusieurs et. VPCs Régions AWS Cela est particulièrement utile pour les applications distribuées qui nécessitent un réseau cohérent et fiable, telles que les microservices, les déploiements entre régions et les environnements cloud natifs complexes.

Amazon VPC Lattice fournit les fonctionnalités clés suivantes :

- Service-to-service mise en réseau — VPC Lattice simplifie la configuration réseau et de sécurité entre les services au sein d'une architecture de microservices. Il fournit une plate-forme unifiée

pour gérer les communications, afin que les services puissent évoluer indépendamment tout en maintenant des performances et une sécurité élevées.

- Réseau inter-VPC — Le réseau VPC est essentiel pour gérer le trafic entre plusieurs régions ou régions. VPCs Il fournit un cadre réseau cohérent qui permet aux services de communiquer de manière fluide, quel que soit leur emplacement physique. Cette fonctionnalité est particulièrement importante pour les applications à grande échelle qui couvrent plusieurs régions VPCs ou régions géographiques.
- Gestion améliorée de la sécurité — En intégrant les politiques de sécurité directement dans la couche réseau, VPC Lattice permet une service-to-service communication à la fois sécurisée et efficace. Cette fonctionnalité réduit la complexité de la gestion de la sécurité dans un environnement distribué, ce qui permet une mise à l'échelle plus facile et une réduction des frais opérationnels.
- Gestion simplifiée du trafic : VPC Lattice propose des fonctionnalités avancées de gestion du trafic, notamment des mécanismes de routage, d'équilibrage de charge et de basculement. Grâce à ces fonctionnalités, le trafic est distribué efficacement entre les services, optimisant les performances du réseau et améliorant l'évolutivité de l'application.

Optimisation des coûts

Pour garantir un contrôle efficace des ressources, la minimisation des coûts de Kubernetes est essentielle pour les entreprises utilisant cette technologie d'orchestration de conteneurs. Il est difficile de suivre correctement les dépenses dans les paramètres Kubernetes en raison de leur complexité, qui inclut de multiples composants tels que des pods et des nœuds. Grâce à l'application de techniques d'optimisation des coûts, les entreprises peuvent voir où leurs ressources sont dépensées et affecter les dépenses de manière appropriée aux départements ou aux projets.

Bien que le dimensionnement dynamique présente des avantages, s'il n'est pas correctement géré, il peut entraîner des dépenses imprévues. Une gestion efficace des coûts permet d'allouer les ressources uniquement lorsqu'elles sont réellement nécessaires, évitant ainsi des augmentations imprévues des dépenses.

Cette section décrit les approches suivantes en matière d'optimisation des coûts :

- Kubecost
- Boucles d'or
- AWS Fargate
- Instances Spot
- Instances réservées
- AWS Instances de Graviton

Kubecost

Kubecost est une solution de gestion des coûts qui aide les entreprises à suivre, contrôler et optimiser leurs dépenses en infrastructure cloud. Il est spécialement conçu pour les clusters Kubernetes. Kubecost vous fournit des informations sur l'utilisation des ressources et une connaissance des coûts en temps réel, ce qui vous permet de mieux comprendre où et dans quelle mesure vos ressources cloud sont utilisées. Grâce à ces informations, vous pouvez optimiser vos dépenses d'infrastructure, améliorer l'efficacité des ressources et prendre des décisions plus éclairées concernant vos investissements dans le cloud.

Kubecost fournit les fonctionnalités clés suivantes :

- Répartition des coûts — Kubecost propose une répartition complète des coûts pour les ressources Kubernetes, y compris les charges de travail, les services, les espaces de noms et les étiquettes. Cette fonctionnalité permet aux équipes de surveiller les coûts par environnement, projet ou équipe.
- Surveillance des coûts en temps réel : elle permet de surveiller en temps réel les coûts du cloud, fournissant aux entreprises un aperçu immédiat des habitudes de dépenses et aidant à prévenir les dépassements de coûts imprévus.
- Recommandations d'optimisation — Kubecost propose des suggestions pratiques pour minimiser l'utilisation des ressources, notamment en réduisant les ressources inactives, en dimensionnant correctement les charges de travail et en maximisant les dépenses de stockage.
- Budgétisation et alertes : les utilisateurs de Kubecost peuvent créer des budgets et recevoir des rappels lorsqu'une dépense approche ou dépasse des critères prédéterminés. Cette fonctionnalité aide les équipes à respecter les contraintes financières.

Boucles d'or

Goldilocks est un utilitaire Kubernetes conçu pour aider les utilisateurs à optimiser leurs demandes de ressources et les limites des charges de travail Kubernetes. Il fournit des recommandations sur la manière de configurer les ressources du processeur et de la mémoire pour les conteneurs exécutés dans un cluster Kubernetes. Ces recommandations vous aident à vous assurer que les applications disposent du bon nombre de ressources pour fonctionner efficacement sans gaspiller. Cette optimisation peut entraîner des économies, une amélioration des performances et une utilisation plus efficace des clusters Kubernetes.

Goldilocks fournit les principales fonctionnalités suivantes :

- Recommandations en matière de ressources — Goldilocks détermine les paramètres idéaux pour les demandes de ressources et les restrictions en analysant les statistiques de consommation de processeur et de mémoire passées pour les charges de travail Kubernetes. Ce faisant, il est plus facile d'éviter le sous-provisionnement ou le surprovisionnement, ce qui peut entraîner des problèmes de performance et un gaspillage de ressources.
- Intégration VPA — Goldilocks utilise le Kubernetes Vertical Pod Autoscaler (VPA) pour collecter des données et fournir des recommandations. Il fonctionne en « mode recommandation », ce qui signifie qu'il ne modifie pas réellement les paramètres des ressources mais fournit des conseils sur ce que devraient être ces paramètres.

- Analyse basée sur les espaces de noms : Goldilocks vous permet de réguler avec précision les charges de travail optimisées et surveillées en vous permettant de cibler des espaces de noms particuliers à des fins d'analyse.
- Tableau de bord visuel — Le tableau de bord Web affiche visuellement les demandes de ressources suggérées et les restrictions, ce qui vous permet de comprendre facilement les données et d'agir en conséquence.
- Fonctionnement non intrusif : Goldilocks ne modifie pas la configuration du cluster car il fonctionne en mode recommandation. Si vous le souhaitez, vous pouvez appliquer manuellement les paramètres de ressources recommandés après avoir examiné les recommandations.

AWS Fargate

Dans le contexte d'Amazon EKS, vous https://docs.aws.amazon.com/eks/latest/userguide/fargate.htmlAWS Fargate permet d'exécuter des pods Kubernetes sans gérer les instances Amazon sous-jacentes. EC2 Il s'agit d'un moteur de calcul sans serveur qui vous permet de vous concentrer sur le déploiement et le dimensionnement d'applications conteneurisées sans vous soucier de l'infrastructure.

AWS Fargate fournit les fonctionnalités clés suivantes :

- Aucune gestion d'infrastructure : Fargate élimine le besoin de provisionner, de gérer ou de dimensionner des instances EC2 Amazon ou des nœuds Kubernetes. AWS gère l'ensemble de la gestion de l'infrastructure, y compris les correctifs et le dimensionnement.
- Isolation au niveau du pod — Contrairement aux nœuds de travail basés sur Amazon, EC2 Fargate fournit une isolation au niveau des tâches ou des pods. Chaque module fonctionne dans son propre environnement informatique isolé, ce qui améliore la sécurité et les performances.
- Mise à l'échelle automatique : Fargate redimensionne automatiquement les pods Kubernetes en fonction de la demande. Vous n'avez pas besoin de gérer les politiques de dimensionnement ou les pools de nœuds.
- Facturation à la seconde : vous ne payez que pour le vCPU et les ressources de mémoire consommées par chaque pod pendant la durée exacte de son exécution, ce qui constitue une option rentable pour certaines charges de travail.
- Réduction des frais généraux : en éliminant le besoin de gérer les EC2 instances, Fargate vous permet de vous concentrer sur le développement et la gestion de vos applications plutôt que sur les opérations d'infrastructure.

Instances Spot

Les instances Spot permettent de réaliser des économies importantes par rapport à la tarification des instances à la demande et constituent une option abordable pour exécuter EC2 des nœuds de travail Amazon dans un cluster Amazon EKS. Cependant, elle AWS peut interrompre les instances Spot dans le cas où une capacité d'instance à la demande est nécessaire. AWS peut récupérer des instances Spot avec un préavis de 2 minutes lorsque la capacité est nécessaire, ce qui les rend moins fiables pour les charges de travail critiques et dynamiques.

Pour les charges de travail sensibles aux coûts et capables de résister aux interruptions, les instances Spot d'Amazon EKS constituent une bonne option. L'utilisation d'une combinaison d'instances ponctuelles et d'instances à la demande dans un cluster Kubernetes vous permet de réaliser des économies sans sacrifier la disponibilité pour des charges de travail vitales.

Les instances Spot fournissent les fonctionnalités clés suivantes :

- Économies de coûts : les instances Spot peuvent être moins coûteuses que la tarification des instances à la demande, ce qui les rend idéales pour les charges de travail sensibles aux coûts.
- Idéal pour les charges de travail tolérantes aux pannes : parfaitement adapté aux charges de travail apatrides et tolérantes aux pannes telles que le traitement par lots, les tâches CI/CD, l'apprentissage automatique ou le traitement de données à grande échelle où les instances peuvent être remplacées sans interruption majeure.
- Intégration de groupes à dimensionnement automatique : Amazon EKS intègre les instances Spot à Kubernetes Cluster Autoscaler, qui peut remplacer automatiquement les nœuds d'instances Spot interrompus par d'autres instances Spot ou instances à la demande disponibles.

Instances réservées

Dans Amazon EKS, les instances réservées sont un modèle de tarification pour les nœuds de EC2 travail Amazon qui exécutent vos charges de travail Kubernetes. En utilisant les instances réservées, vous vous engagez à utiliser des types d'instances spécifiques pour une durée d'un ou trois ans, en échange d'économies par rapport à la tarification des instances à la demande. La réservation d'instances dans Amazon EKS est un moyen abordable d'effectuer des charges de travail cohérentes et à long terme sur les nœuds de travail Amazon EC2 .

Les instances réservées sont couramment utilisées pour Amazon EC2. Toutefois, les nœuds de travail de votre cluster Amazon EKS (qui sont des EC2 instances) peuvent également bénéficier de

ce modèle économique, à condition que la charge de travail nécessite une utilisation prévisible et à long terme.

Les services de production, les bases de données et les autres applications dynamiques qui nécessitent une haute disponibilité et des performances constantes sont des exemples de charges de travail stables parfaitement adaptées aux instances réservées.

Les instances réservées fournissent les fonctionnalités clés suivantes :

- Économies de coûts : les instances réservées permettent de réaliser des économies par rapport aux instances à la demande, en fonction de la durée du contrat (1 ou 3 ans) et du plan de paiement (paiement initial intégral, paiement initial partiel ou absence de paiement initial).
- Engagement à long terme : vous vous engagez pour une durée d'un ou trois ans pour un type, une taille et une taille d'instance spécifiques. Région AWS C'est idéal pour les charges de travail stables et exécutées en continu dans le temps.
- Tarification prévisible : dans la mesure où vous vous engagez à respecter une durée spécifique, les instances réservées fournissent des coûts mensuels ou initiaux prévisibles, ce qui facilite la budgétisation des charges de travail à long terme.
- Flexibilité des instances : avec les instances réservées convertibles, vous pouvez modifier le type, la famille ou la taille de l'instance pendant la période de réservation. Les instances réservées convertibles offrent plus de flexibilité que les instances réservées standard, qui n'autorisent pas les modifications.
- Capacité garantie : les instances réservées garantissent la disponibilité de la capacité dans la zone de disponibilité où la réservation est effectuée, ce qui est crucial pour les charges de travail critiques nécessitant une puissance de calcul constante.
- Aucun risque d'interruption — Contrairement aux instances ponctuelles, les instances réservées ne sont pas susceptibles d'être interrompues par AWS. Elles sont donc idéales pour exécuter des charges de travail critiques qui nécessitent une disponibilité garantie.

AWS Instances de Graviton

AWS Graviton est une famille de processeurs ARM conçus pour améliorer les performances et AWS la rentabilité des charges de travail dans le cloud. Dans le contexte d'Amazon EKS, vous pouvez utiliser des instances Graviton comme nœuds de travail pour exécuter vos charges de travail Kubernetes, ce qui vous permet de réaliser des gains de performances et des économies de coûts significatifs.

Les instances Graviton constituent une excellente option pour les applications natives du cloud et gourmandes en ressources informatiques, car elles offrent un rapport qualité-prix supérieur à celui des instances x86. Toutefois, lorsque vous envisagez d'adopter des instances Graviton, tenez compte de la compatibilité ARM.

AWS Les instances Graviton fournissent les fonctionnalités clés suivantes :

- Architecture basée sur ARM — Les processeurs AWS Graviton sont basés sur l'architecture ARM, qui est différente des architectures x86 traditionnelles mais très efficace pour de nombreuses charges de travail.
- Rentable : les EC2 instances Amazon basées sur Graviton offrent généralement un meilleur rapport prix/performances par rapport aux instances x86. EC2 Cela en fait une option intéressante pour les clusters Kubernetes qui exécutent Amazon EKS.
- Performances — Les processeurs Graviton2, la deuxième génération de AWS Graviton, offrent des améliorations significatives en termes de performances de calcul, de débit de mémoire et d'efficacité énergétique. Elles sont idéales pour les charges de travail gourmandes en CPU et en mémoire.
- Différents types d'instances : les instances Graviton se déclinent en différentes familles, telles que t4g, m7g, c7g et r7g, couvrant une gamme de cas d'utilisation allant des charges de travail à usage général aux charges de travail optimisées pour le calcul, optimisées pour la mémoire et évolutives.
- Groupes de nœuds Amazon EKS : vous pouvez configurer des groupes de nœuds gérés par Amazon EKS ou des groupes de nœuds autogérés pour inclure des instances basées sur Graviton. Grâce à cette approche, vous pouvez exécuter des charges de travail optimisées pour l'architecture ARM sur le même cluster Kubernetes avec des instances x86.

Étapes suivantes

Ce guide fournit des informations pour vous aider à optimiser Amazon EKS en termes de dimensionnement du calcul, de dimensionnement de la charge de travail, de dimensionnement du réseau et d'optimisation des coûts. En comprenant et en appliquant ces concepts, les entreprises peuvent créer un environnement cloud hautement efficace, évolutif et rentable qui répond à leurs besoins dynamiques.

La mise en œuvre efficace du calcul et de la mise à l'échelle de la charge de travail permet de garantir que les ressources sont utilisées efficacement et que les applications maintiennent des performances élevées même pendant les périodes de pointe. L'adoption de techniques de mise à l'échelle du réseau, telles que la mise en réseau personnalisée et la délégation de préfixes, favorise la gestion des ressources du réseau et une évolutivité sans faille. Mettre l'accent sur l'optimisation des coûts aide les organisations à trouver un équilibre entre performance et efficacité financière.

L'intégration de ces conseils dans votre stratégie cloud peut vous aider à améliorer les performances et l'évolutivité de votre infrastructure et à réaliser des économies. Cette approche globale peut vous permettre de créer un environnement cloud robuste qui soutient la croissance de votre entreprise et s'adapte aux exigences commerciales en constante évolution.

Ressources

AWS blogues

- Favoriser l'optimisation des coûts et la résilience pour EKS avec des instances ponctuelles
- Combiner AWS Graviton et x86 CPUs pour optimiser les coûts et la résilience à l'aide d'Amazon EKS

AWS documentation

- CNI Amazon VPC
- Amazon Elastic Kubernetes Service AWS (livre blanc : Présentation des options de déploiement sur) AWS
- Guide des meilleures pratiques Amazon EKS
- Charpentier
- En savoir plus sur Kubecost
- Simplifiez la gestion du calcul avec AWS Fargate

Autres ressources

- Mise à l'échelle automatique du cluster (documentation Kubernetes)
- Goldilocks : un outil open source pour recommander des demandes de ressources (Fairwinds Blog)
- Mise à l'échelle automatique des modules horizontaux (documentation Kubernetes)
- Kubecost (documentation Kubecost)
- Autoscaling piloté par les événements de Kubernetes (documentation KEDA)

Historique du document

Le tableau suivant décrit les modifications importantes apportées à ce guide, intitulé Mise à l'échelle de l'infrastructure Amazon EKS pour optimiser le calcul, les charges de travail et les performances du réseau. Pour être averti des mises à jour à venir, abonnez-vous à un fil RSS.

Modification	Description	Date
Publication initiale	—	11 novembre 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 :

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

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

A

A2 (1) Agent-to-Agent

Protocole dynamique pour la collaboration agent-agent prenant en charge la délégation de tâches et le transfert d'état.

ABAC

Voir contrôle d'accès basé sur les attributs.

services abstraits

Consultez la section Services gérés.

ACIDE

Voir atomicité, consistance, isolation, durabilité.

migration active-active

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

migration active-passive

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

Agent

Un système d'IA capable de raisonner, de planifier et de prendre des mesures de manière autonome à l'aide d'outils pour atteindre des objectifs.

Agent Ops

Pratiques opérationnelles pour la création, le test, le déploiement et l'exécution d'agents d'IA en production à grande échelle.

fonction d'agrégation

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

AI

Voir intelligence artificielle.

AIOps

Voir les opérations d'intelligence artificielle.

anonymisation

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

anti-motif

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

contrôle des applications

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

portefeuille d'applications

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

intelligence artificielle (IA)

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

opérations d'intelligence artificielle (AIOps)

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

chiffrement asymétrique

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

atomicité, cohérence, isolement, durabilité (ACID)

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

contrôle d'accès par attributs (ABAC)

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

source de données faisant autorité

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

Zone de disponibilité

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

AWS Cadre d'adoption du cloud (AWS CAF)

Un cadre de directives et de meilleures pratiques visant AWS à aider les entreprises à élaborer un plan efficace pour réussir leur migration vers le cloud. AWS La CAF organise ses conseils en six domaines prioritaires appelés perspectives : les affaires, les personnes, la gouvernance, les plateformes, la sécurité et les opérations. Les perspectives d'entreprise, de personnes et de gouvernance mettent l'accent sur les compétences et les processus métier, tandis que les

perspectives relatives à la plateforme, à la sécurité et aux opérations se concentrent sur les compétences et les processus techniques. Par exemple, la perspective liée aux personnes cible les parties prenantes qui s'occupent des ressources humaines (RH), des fonctions de dotation en personnel et de la gestion des personnes. Dans cette perspective, la AWS CAF fournit des conseils pour le développement du personnel, la formation et les communications afin de préparer l'organisation à une adoption réussie du cloud. Pour plus d'informations, veuillez consulter le site Web AWS CAF et le livre blanc AWS CAF.

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

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

B

mauvais bot

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

BCP

Consultez la section Planification de la continuité des activités.

graphique de comportement

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

système de poids fort

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

classification binaire

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

filtre de Bloom

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

blue/green déploiement

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 Mettre en œuvre des procédures permettant de briser le verre dans le AWS Well-Architected guide.

stratégie existante (brownfield)

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

cache de tampon

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

capacité métier

Ce que fait une entreprise pour générer de la valeur (par exemple, les ventes, le service client ou le marketing). Les architectures de microservices et les décisions de développement peuvent être dictées par les capacités métier. Pour plus d'informations, veuillez consulter la section Organisation en fonction des capacités métier du livre blanc Exécution de microservices conteneurisés sur AWS.

planification de la continuité des activités (BCP)

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

C

CAF

Voir le cadre d'adoption du AWS cloud.

déploiement de Canary

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

CCoE

Voir le Centre d'excellence du cloud.

CDC

Consultez la section Capture des données de modification.

capture des données de modification (CDC)

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

ingénierie du chaos

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

CI/CD

Découvrez l'intégration continue et la livraison continue.

classification

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

Développeur citoyen

Un utilisateur professionnel qui crée des applications d'intelligence artificielle à l'aide de plateformes sans code/low code sans compétences techniques spécialisées.

chiffrement côté client

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

Centre d'excellence cloud (CCoE)

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

cloud computing

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

modèle d'exploitation du cloud

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

étapes d'adoption du cloud

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

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

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

CMDB

Consultez la base de données de gestion des configurations.

référentiel de code

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

cache passif

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

données gelées

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

vision par ordinateur (CV)

Domaine de l'IA qui utilise l'apprentissage automatique pour analyser et extraire des informations à partir de formats visuels tels que des images numériques et des vidéos. Par exemple, Amazon SageMaker AI fournit des algorithmes de traitement d'image pour les CV.

dérive de configuration

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

base de données de gestion des configurations (CMDB)

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

pack de conformité

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

intégration continue et livraison continue (CI/CD)

Processus d'automatisation des étapes de source, de construction, de test, de préparation et de production du processus de publication du logiciel. CI/CD est communément décrit comme un pipeline. CI/CD peut vous aider à automatiser les processus, à améliorer la productivité,

à améliorer la qualité du code et à accélérer les livraisons. Pour plus d'informations, veuillez consulter Avantages de la livraison continue. CD peut également signifier déploiement continu. Pour plus d'informations, veuillez consulter Livraison continue et déploiement continu.

CV

Voir vision par ordinateur.

D

données au repos

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

classification des données

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

défense en profondeur

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 approche de défense approfondie 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 la section Reprise après sinistre des charges de travail sur AWS : Restauration dans le cloud dans le 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 livre, Domain-Driven Design : Tackling Complexity in the Heart of Software (Boston : Addison-Wesley Professional, 2003). Pour plus d'informations sur la manière dont vous pouvez utiliser la conception axée sur le domaine avec le modèle Strangler Fig, consultez la section Modernisation incrémentielle des anciens services Web ASP.NET Microsoft (ASMX) à l'aide de conteneurs et d'Amazon API Gateway.

DR

Consultez la section Reprise après sinistre.

détection de dérive

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

DVSM

Voir la cartographie de la chaîne de valeur du développement.

E

EDA

Voir analyse exploratoire des données.

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 dans lequel les octets sont stockés dans la mémoire de l'ordinateur. Big-endian les systèmes stockent d'abord l'octet le plus significatif. Little-endian les systèmes stockent d'abord l'octet le moins significatif.

point de terminaison

Voir point de terminaison de service.

service de point de terminaison

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

planification des ressources d'entreprise (ERP)

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

chiffrement d'enveloppe

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

environment

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

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

épopée

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

ERP

Voir Planification des ressources d'entreprise.

analyse exploratoire des données (EDA)

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

F

tableau des faits

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

échouer rapidement

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

limite d'isolation des défauts

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

branche de fonctionnalités

Voir la succursale.

fonctionnalités

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

importance des fonctionnalités

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

transformation de fonctionnalité

Optimiser les données pour le processus de ML, notamment en enrichissant les données avec des sources supplémentaires, en mettant à l'échelle les valeurs ou en extrayant plusieurs ensembles d'informations à partir d'un seul champ de données. Cela permet au modèle de ML

de tirer parti des données. Par exemple, si vous décomposez la date « 2021-05-27 00:15:37 » en « 2021 », « mai », « jeudi » et « 15 », vous pouvez aider l'algorithme d'apprentissage à apprendre des modèles nuancés associés à différents composants de données.

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. Few-shot l'envoi d'instructions peut être efficace pour les tâches qui nécessitent un formatage, un raisonnement ou une connaissance du 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'entraîne sur des ensembles de données massifs de données généralisées et non étiquetées. Les FM sont capables d'effectuer une grande variété de tâches générales, telles que la compréhension du langage, la génération de texte et d'images et la conversation en langage naturel. Pour plus d'informations, voir Que sont les modèles de base ?

Passerelle FM

Un intermédiaire centralisé qui contrôle et normalise l'accès aux modèles de base. Également connue sous le nom de passerelle LLM.

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

rambardes (AI)

Des mécanismes de sécurité qui filtrent, valident et limitent les entrées et sorties des agents afin de garantir un comportement responsable et sûr de l'IA.

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.

humain dans la boucle (HiTL)

Un modèle de flux de travail dans lequel l'exécution des agents s'arrête pour examen et approbation par l'homme aux points de décision critiques.

migration de base de données homogène

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

données chaudes

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

correctif

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

période de soins intensifs

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

I

IaC

Considérez l'infrastructure comme un code.

politique basée sur l’identité

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

application inactive

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

IIoT

Voir Internet industriel des objets.

infrastructure immuable

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

VPC entrant (d'entrée)

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

migration incrémentielle

Stratégie de basculement dans le cadre de laquelle vous migrez votre application par petites parties au lieu d'effectuer un basculement complet unique. Par exemple, il se peut que vous ne transfériez que quelques microservices ou utilisateurs vers le nouveau système dans un

premier temps. Après avoir vérifié que tout fonctionne correctement, vous pouvez transférer progressivement des microservices ou des utilisateurs supplémentaires jusqu'à ce que vous puissiez mettre hors service votre système hérité. Cette stratégie réduit les risques associés aux migrations de grande ampleur.

Industry 4.0

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

infrastructure

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

infrastructure en tant que code (IaC)

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

internet industriel des objets (IIoT)

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

VPC d'inspection

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

Internet des objets (IoT)

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

interprétabilité

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

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 la succursale.

malware

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

services gérés

Services AWS 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.

MCP

Voir Model Context Protocol.

Protocole de contexte du modèle (MCP)

Protocole sans état pour la communication entre un agent et un outil.

serveur MCP

Service qui expose un ou plusieurs outils via le protocole Model Context.

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 AWS Well-Architected cadre.

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)

Un protocole de communication léger de machine à machine (M2M), basé sur le publish/subscribemodèle, pour les appareils IoT aux ressources limitées.

microservice

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

architecture de microservices

Approche de création d'une application avec des composants indépendants qui exécutent chaque processus d'application en tant que microservice. Ces microservices communiquent via une interface bien définie à l'aide d'API légères. Chaque microservice de cette architecture peut être mis à jour, déployé et mis à l'échelle pour répondre à la demande de fonctions spécifiques d'une application. Pour plus d'informations, consultez la section Implémentation de microservices sur AWS.

Programme d'accélération des migrations (MAP)

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

migration à grande échelle

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

usine de migration

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

métadonnées de migration

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

modèle de migration

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

Évaluation du portefeuille de migration (MPA)

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

Évaluation de la préparation à la migration (MRA)

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

stratégie de migration

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

ML

Voir apprentissage automatique.

modernisation

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

évaluation de la préparation à la modernisation

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

applications monolithiques (monolithes)

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

MPA

Voir Évaluation du portefeuille de migration.

MQTT

Voir Message Queuing Telemetry Transport.

classification multi-classes

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

infrastructure mutable

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

O

OAC

Voir Contrôle d'accès à l'origine.

OAI

Voir l'identité d'accès à l'origine.

OCM

Voir gestion du changement organisationnel.

migration hors ligne

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

OI

Consultez la section Intégration des opérations.

OLA

Voir l'accord au niveau opérationnel.

migration en ligne

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

OPC-UA

Voir Open Process Communications - Architecture unifiée.

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

Protocole de communication machine à machine (M2M) pour l'automatisation industrielle. OPC-UA fournit une norme d'interopérabilité avec des schémas de chiffrement, d'authentification et d'autorisation des données.

accord au niveau opérationnel (OLA)

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

examen de l'état de préparation opérationnelle (ORR)

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

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). OAC prend en charge tous les compartiments S3 dans leur ensemble Régions AWS, le chiffrement côté serveur avec AWS KMS (SSE-KMS) et les `DELETE` requêtes dynamiques `PUT` adressées au compartiment S3.

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 de référence de sécuritéAWS recommande de configurer votre compte réseau avec des VPC entrants, sortants et d'inspection afin de protéger l'interface bidirectionnelle entre votre application et Internet en général.

P

limite des autorisations

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

informations personnelles identifiables (PII)

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

PII

Voir les informations personnelles identifiables.

manuel stratégique

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

PLC

Voir contrôleur logique programmable.

PLM

Consultez la section Gestion du cycle de vie des produits.

policy

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

persistance polyglotte

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

évaluation du portefeuille

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

predicate

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

prédicat pushdown

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

contrôle préventif

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

principal

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

contrôle proactif

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

gestion du cycle de vie des produits (PLM)

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

environnement de production

Voir environnement.

contrôleur logique programmable (PLC)

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

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).

RAG

Voir Retrieval Augmented Generation.

rançongiciel

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

Matrice RASCI

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

RCAC

Voir contrôle d'accès aux lignes et aux colonnes.

réplica en lecture

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

réarchitecte

Voir 7 Rs.

objectif de point de récupération (RPO)

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

objectif de temps de récupération (RTO)

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

refactoriser

Voir 7 Rs.

Région

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

régression

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

réhéberger

Voir 7 Rs.

version

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

déplacer

Voir 7 Rs.

replateforme

Voir 7 Rs.

rachat

Voir 7 Rs.

résilience

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

runbook

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

S

SAML 2.0

Un standard ouvert utilisé par de nombreux fournisseurs d'identité (IdPs). Cette fonctionnalité permet l'authentification unique fédérée (SSO), afin que les utilisateurs puissent se connecter

AWS Management Console ou appeler les opérations de l' AWS API sans que vous ayez à créer un utilisateur dans IAM pour tous les membres de votre organisation. Pour plus d'informations sur la fédération SAML 2.0, veuillez consulter À propos de la fédération SAML 2.0 dans la documentation IAM.

SCADA

Voir Contrôle de supervision et acquisition de données.

SCP

Voir la politique de contrôle des services.

secret

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

sécurité dès la conception

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

contrôle de sécurité

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

renforcement de la sécurité

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

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

Outils et services qui associent les systèmes de gestion des informations de sécurité (SIM) et de gestion des événements de sécurité (SEM). Un système SIEM collecte, surveille et analyse les

données provenant de serveurs, de réseaux, d'appareils et d'autres sources afin de détecter les menaces et les failles de sécurité, mais aussi de générer des alertes.

automatisation des réponses de sécurité

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

chiffrement côté serveur

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

Politique de contrôle des services (SCP)

Politique qui propose un contrôle centralisé des autorisations pour tous les comptes d'une organisation dans AWS Organizations. Les SCP définissent des barrières de protection ou des limites aux actions qu'un administrateur peut déléguer à des utilisateurs ou à des rôles. Vous pouvez utiliser les SCP comme listes d'autorisation ou de refus, pour indiquer les services ou les actions autorisés ou interdits. Pour plus d'informations, consultez la section Politiques de contrôle des services dans la AWS Organizations documentation.

point de terminaison du service

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

contrat de niveau de service (SLA)

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

indicateur de niveau de service (SLI)

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

objectif de niveau de service (SLO)

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

modèle de responsabilité partagée

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

IA de l'ombre

Applications d'IA non autorisées créées ou utilisées en dehors des canaux régis au sein d'une organisation.

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.

modèle split-and-seed

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 un exemple d'application de ce modèle, consultez la section Modernisation progressive des anciens services Web Microsoft ASP.NET (ASMX) à l'aide de conteneurs et d'Amazon API Gateway.

sous-réseau

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

contrôle de supervision et acquisition de données (SCADA)

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

chiffrement symétrique

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

tests synthétiques

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

invite du système

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

T

tags

Key-value des paires 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.

outil

Fonction ou API qu'un agent peut invoquer pour effectuer des opérations dans des systèmes externes.

passerelle de transit

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

flux de travail basé sur jonction

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

accès sécurisé

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

réglage

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

équipe de deux pizzas

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

U

incertitude

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

tâches indifférenciées

Également connu sous le nom de « levage de charges lourdes », ce travail est nécessaire pour créer et exploiter une application, mais qui n'apporte pas de valeur directe à l'utilisateur final ni

d'avantage concurrentiel. Les exemples de tâches indifférenciées incluent l'approvisionnement, la maintenance et la planification des capacités.

environnements supérieurs

Voir environnement.

V

mise à vide

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

contrôle de version

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

Appairage de VPC

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

vulnérabilités

Défaut logiciel ou matériel qui compromet la sécurité du système.

W

cache actif

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

données chaudes

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

fonction de fenêtre

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

charge de travail

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

flux de travail

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

VER

Voir écrire une fois, lire plusieurs.

WQF

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 de type « jour zéro »

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

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.