Économie pour l'IA agentique sur AWS

AWS Directives prescriptives

AWS Directives prescriptives: Économie pour l'IA agentique sur AWS

Table of Contents

Économie pour l'IA agentique sur AWS

Hans Schabert et Prasanta Roy, Amazon Web Services

Janvier 2026 (historique du document)

Organisations adoptant des systèmes d'automatisation pilotés par l'IA et d'IA agentique doivent prendre des décisions économiques éclairées entre le travail humain et les agents intelligents. Cela devient essentiel pour des opérations cloud durables. Ce guide vous aide à évaluer, à mettre en œuvre et à optimiser les compromis économiques entre la main-d'œuvre humaine et les systèmes d'IA agentic sur. AWS Vous pouvez optimiser votre retour sur investissement (ROI) tout en préservant l'excellence opérationnelle.

Aucun système n'est correct à 100 %. Ce principe fondamental sous-tend l'analyse économique des systèmes d'IA humains et agentiques. Organisations doivent aller au-delà des comparaisons de coûts simplistes pour évaluer l'impact économique total, les profils de risque, les exigences de qualité des décisions et la création de valeur stratégique à long terme.

Le comportement des clients évolue radicalement, passant des investissements technologiques initiaux traditionnels à des pay-per-outcome modèles qui alignent les coûts sur les résultats commerciaux. Cette transformation nécessite de nouvelles approches pour l'évaluation, la mise en œuvre et l'optimisation de la collaboration homme-agent.

Le chemin vers le succès suit un schéma clair : commencez par les emplois appropriés, mesurez tout et adaptez ce qui fonctionne à grande échelle. Organisations qui adoptent cette approche obtiennent un avantage concurrentiel durable grâce à une allocation intelligente des ressources et à une automatisation axée sur les résultats.

Public visé

Ce guide est destiné aux personnes suivantes :

- Dirigeants (CEOs CTOs,, CFOs) qui prennent des décisions d'investissement stratégiques
- Architectes d'entreprise qui conçoivent des stratégies d'automatisation organisationnelle
- Praticiens des opérations financières qui optimisent la gestion financière dans le cloud
- Les leaders technologiques qui évaluent les approches de mise en œuvre de l'IA
- Dirigeants d'unités commerciales qui souhaitent comprendre le retour sur investissement de l'automatisation

- Les professionnels de l'approvisionnement qui utilisent les nouveaux modèles de tarification basés sur l'IA

Pour comprendre les concepts de ce guide, nous vous recommandons de consulter les fondements de l'IA agentic sur. AWS

Objectifs

Ce guide vous aide à comprendre les points suivants :

- Comment évaluer les emplois en fonction du potentiel d'automatisation agentique
- Modèles économiques pour comparer les coûts du travail humain aux investissements dans les systèmes d'IA agentiques
- Pay-per-outcome les modèles de tarification et leur impact sur l'économie des projets d'IA
- Techniques de mesure pour démontrer le retour sur investissement et gérer les risques
- Stratégies de mise à l'échelle qui transforment les coûts fixes en résultats variables

À propos de cette série de contenus

Ce guide fait partie d'une série sur l'IA agentique sur. AWS Pour plus d'informations et pour consulter les autres guides de cette série, consultez Agentic AI sur le site Web de AWS Prescriptive Guidance.

Comprendre l'économie de l'IA agentique sur AWS

L'un des principes clés consiste à déterminer quand utiliser des agents d'IA et quand utiliser des méthodes déterministes traditionnelles. Organisations doivent évaluer systématiquement les emplois qui justifient une automatisation agentique et ceux qui devraient recourir à l'automatisation traditionnelle ou à une exploitation humaine continue. Cette décision nécessite de comprendre la relation entre les caractéristiques de la tâche, la tolérance au risque et l'approche opérationnelle.

Avant de décider de mettre en œuvre l'IA agentic, vous devez utiliser le cadre décisionnel pour comprendre l'impact économique. Le cadre décisionnel comprend les trois questions clés suivantes :

1. Évaluation des tâches — Cette tâche convient-elle à un agent d'intelligence artificielle ?
2. Évaluation de l'impact des risques — Quels sont les risques encourus ?
3. Retour sur investissement — Sera-t-il rentable ?

Évaluation des tâches

Les tâches comportant des règles de décision standardisées très complexes peuvent bénéficier des approches de l'IA agentique. Les tâches simples et hautement standardisées sont mieux servies par l'automatisation traditionnelle ou l'automatisation robotique des processus. Les systèmes d'IA agentic excellent dans le raisonnement, la compréhension du contexte ou la prise de décisions adaptative. Ils ajoutent de la valeur au-delà du traitement basé sur des règles. Les mises en œuvre réussies de l'IA agentique nécessitent des systèmes capables d'apprendre et de s'adapter.

Tenez compte des facteurs suivants lors de l'évaluation d'une tâche :

- Complexité — Degré de raisonnement et de compréhension du contexte requis. Les tâches nécessitant une compréhension contextuelle, une interprétation nuancée ou des réponses adaptatives aux conditions changeantes favorisent les approches agentiques par rapport à l'automatisation traditionnelle, tandis que les tâches purement mécaniques ou de calcul peuvent ne pas nécessiter d'intelligence agentique.
- Normalisation — Présence de modèles et de règles clairs. L'IA agentic est recommandée si la tâche nécessite une compréhension contextuelle. Si aucune adaptation ou apprentissage n'est nécessaire, pensez à l'automatisation traditionnelle.
- Volume — Fréquence d'exécution des tâches. L'IA agentic est recommandée pour les activités autonomes. L'automatisation traditionnelle est recommandée pour les tâches cohérentes et

volumineuses. Cependant, le volume à lui seul ne détermine pas l'approche. Les décisions de faible volume et de grande valeur peuvent justifier l'assistance d'un agent pour améliorer la qualité des décisions plutôt que pour réduire les coûts.

- Valeur — Impact commercial par tâche accomplie. Envisagez l'IA agentique pour obtenir des résultats de grande valeur qui nécessitent une capacité autonome semblable à celle de l'homme. Envisagez l'automatisation traditionnelle pour les tâches répétitives et cohérentes, qui peuvent être effectuées de manière déterministe.

Évaluation de l'impact des risques

Il existe actuellement quatre approches de déploiement de l'IA agentique : totalement autonome, humaine intégrée, copilote ou dirigée par un humain avec le soutien d'un agent. Chacun a son propre profil de risque et sa propre tolérance aux erreurs, et tous impliquent des humains d'une manière ou d'une autre. Le tableau suivant décrit les détails des risques liés à ces approches.

Niveau d'autonomie	Profil de risque	Tolérance aux erreurs	Exemples de cas d'utilisation	Engagement humain
Totalement autonome	Risque faible	1 à 2 % acceptables	• Catégorisation des données de base • Acheminement des documents • Génération de rapports standard	• Supervision minimale • Audits périodiques
Un humain au courant	Risque moyen	Inférieur à 0,5 %	• Projets de réponses • Modération du contenu	• Révision régulière • Gestion des exceptions

			• Traitement initial des demandes	• Assurance qualité
Copilote	Risque élevé	Près de zéro	• Contribution à la planification stratégique • Évaluation des risques • Décisions d'investissement	• L'humain prend les décisions finales • L'agent fournit des recommandations
Dirigé par l'homme avec le soutien d'agents	Risque critique	Tolérance zéro	• Décisions juridiques • Diagnostic médical • Conformité aux réglementations	• Processus lié aux pulsions humaines • L'agent fournit des recherches ou des analyses et des informations complémentaires uniquement

Le tableau suivant décrit les principales considérations à prendre en compte lors du choix entre ces approches.

Considération	Totalement autonome	Un humain au courant	Copilote	Dirigé par des

Cost efficiency (Rentabilité)	Le plus élevé	Élevée	Moyenne	Faible
Capacité de mise à l’échelle	Illimité	Élevée	Moyenne	Limité
Vitesse de traitement	Le plus rapide	Rapide	Moyenne	Lent
Gestion des risques	Base	Amélioré	Solide	Le plus fort
Gestion de la complexité	Tâches simples	Tâches modérément complexes	Tâches complexes	Tâches critiques

Ce cadre de prise en compte aide les organisations à adapter les niveaux d'autonomie aux profils de risque, à adapter les opérations de manière appropriée, à équilibrer efficacité et contrôle, à mettre en œuvre une gouvernance appropriée et à optimiser l'allocation des ressources.

Retour sur investissement

Le calcul du retour sur investissement des systèmes d'intelligence artificielle agentic commence par une analyse complète des coûts. Organisations doivent d'abord calculer leurs coûts humains actuels, y compris les salaires, les avantages sociaux et les dépenses liées à l'espace de travail, ainsi que les dépenses spécifiques aux processus et les coûts cachés tels que la formation, la couverture et les temps d'arrêt.

Pour l'analyse du seuil de rentabilité, les entreprises doivent tenir compte des coûts de mise en œuvre, des dépenses opérationnelles permanentes et du volume nécessaire pour justifier l'investissement. Il est également important de tenir compte des variations saisonnières et des avantages liés à la courbe d'apprentissage qui apparaissent à mesure que les systèmes mûrissent et s'améliorent au fil du temps.

Lors de l'évaluation des agents d'intelligence artificielle, les entreprises doivent se rappeler que ces systèmes ont généralement des coûts initiaux plus élevés mais des coûts par transaction inférieurs à ceux des opérations humaines. De plus, les agents d'intelligence artificielle améliorent leurs performances au fil du temps et offrent une meilleure évolutivité que les équipes humaines. Cela les

rend de plus en plus rentables à mesure que le déploiement prend de l'ampleur et que l'expérience opérationnelle s'accumule.

Mesurer le succès et le retour sur investissement des systèmes d'IA agentic

Mesurer le succès de la mise en œuvre d'un système d'IA agentique nécessite une approche systématique. Cette section fournit une méthodologie claire pour l'évaluation et l'optimisation continue qui utilise votre analyse existante plutôt que de partir de zéro.

Étape 1 : Utilisez votre fondation existante

Commencez par une évaluation complète des coûts conformément aux recommandations de la section Évaluation des coûts de vos processus actuels. Cela fournit une base opérationnelle pour vos calculs de retour sur investissement. Comme décrit dans la section Évaluation de l'impact des risques, choisissez entre les quatre niveaux d'autonomie (entièrement autonome, boucle humaine, approche copilote, dirigée par un humain avec le soutien d'un agent) afin de déterminer les critères de mesure appropriés et les seuils de tolérance aux erreurs pour chaque processus.

Étape 2 : Définissez des objectifs de réussite clairs

Établissez une architecture et des objectifs de réussite qui mettent l'accent sur les systèmes capables d'apprentissage, comme décrit dans la section Modèles de réussite pour la mise en œuvre de systèmes d'IA agentiques. Concentrez-vous sur l'amélioration continue plutôt que sur les performances statiques. Définissez des délais de retour sur investissement en utilisant la méthodologie d'analyse du seuil de rentabilité démontrée dans l'étude de cas : comparaison des coûts de l'IA humaine et agentique pour les opérations de recrutement. Incluez des points de décision clairs pour mettre fin aux agents non performants.

Étape 3 : suivre les indicateurs clés

Surveillez les performances financières par rapport à votre base de référence établie, et suivez les économies de coûts et les améliorations de valeur stratégiques. Mesurez les indicateurs opérationnels, notamment les taux d'erreur dans les limites acceptables pour le niveau d'autonomie que vous avez choisi, les améliorations de la vitesse de traitement et les gains de cohérence. Concentrez-vous sur les indicateurs stratégiques qui démontrent la capacité d'apprentissage et l'adaptation au fil du temps.

Étape 4 : Utilisation AgentOps

Appliquez le cadre d'apprentissage continu de la section Intégration du feedback humain dans les systèmes d'IA agentic pour optimiser la prise de décision grâce à l'intégration systématique du feedback humain. Créez des systèmes d'apprentissage en temps réel qui intègrent les connaissances humaines pour améliorer les performances. Surveillez la transformation vers des modèles commerciaux basés sur les résultats, comme décrit dans Transformation économique vers une tarification basée sur les résultats pour les systèmes d'IA agentic sur. AWS

Évaluation de vos coûts actuels liés aux processus humains

Comprendre les coûts réels de vos processus est essentiel pour prendre des décisions éclairées concernant les investissements dans les systèmes d'IA agentic. Tout d'abord, vous devez établir une base de référence précise du coût de vos processus actuels, y compris toutes les dépenses cachées, les taux d'échec et les coûts d'opportunité. Cela vous permet de calculer le retour sur investissement avec précision et de prendre des décisions stratégiques. Cette évaluation complète des coûts constitue la base essentielle pour déterminer si les systèmes d'IA agentic peuvent apporter une véritable valeur ajoutée en tant que compagnons de productivité.

L'évaluation des coûts de référence est essentielle pour les principales raisons suivantes :

- Précision du retour sur investissement — Des bases de coûts précises permettent de réaliser des projections de retour sur investissement réalistes qui tiennent compte de l'ensemble des dépenses opérationnelles actuelles.
- Stratégie de mise en œuvre des agents : une compréhension complète des coûts aide les entreprises à identifier les processus les plus prometteurs pour le déploiement initial d'un système d'IA agentique.
- Mesure du rendement — Les bases de référence établies fournissent le cadre de mesure permettant de suivre les avantages réels et prévus des mises en œuvre de l'IA agentique.

Organisations doivent systématiquement identifier et évaluer tous les facteurs de coût qui influencent l'économie des processus avant de comparer les alternatives humaines et agentiques. Cette évaluation garantit des calculs de référence précis en tenant compte des facteurs de coûts évidents et cachés. Il met particulièrement l'accent sur les coûts d'échec, les taux d'échec historiques et les opportunités commerciales manquées, qui représentent le véritable coût total des processus actuels.

Cette section décrit comment collecter des données dans chaque catégorie de coûts afin d'établir des mesures de référence précises pour vos processus actuels. Il décrit les sources d'information et fournit des exemples pour les catégories de coûts suivantes :

Coûts de main-d'œuvre

Extrayez 24 mois de données salariales comprenant le salaire de base, les heures supplémentaires, les avantages sociaux et les coûts de formation. Utilisez votre système d'information sur les ressources humaines (SIRH) pour suivre les dépenses de recrutement et les taux de rotation. Les systèmes de suivi du temps révèlent la productivité réelle par rapport aux heures planifiées. Les plateformes de gestion des performances montrent la corrélation entre les niveaux de compétence et les coûts de rémunération. Calculez les taux horaires complets qui sont alloués aux frais de gestion.

Voici un exemple de liste d'inducteurs de coûts liés à la main-d'œuvre.

Facteur de coûts	Impact commercial
Rémunération de base	25 à 150$ par heure à pleine charge
Avantages sociaux et charges sociales	25 à 40 % du salaire de base
Formation et développement	5 à 15 % du coût annuel de la main-d'œuvre
Frais de gestion	15 à 25 % du coût direct de la main-d'œuvre

Coûts liés à la performance humaine et à la cohérence

Combinez les données des systèmes de gestion de projet qui indiquent les variations de réalisation des tâches avec les systèmes de présence. Cela peut révéler des tendances en matière d'absentéisme et des changements saisonniers. Les plateformes de service client présentent les plages de performance individuelles grâce à des indicateurs de résolution, et les données de gestion

de la relation client (CRM) des ventes peuvent indiquer les variations d'efficacité lors de la conclusion des transactions. Les systèmes de gestion de la qualité fournissent les taux de défauts et traitent les données de conformité entre les équipes et les sites. Les systèmes de flux de travail enregistrent les délais d'exécution, les délais d'approbation et la fréquence de traitement des exceptions. L'analyse des communications révèle les coûts de coordination liés à la fréquence des réunions et aux modèles de collaboration.

Voici une liste d'exemples de facteurs de coûts liés à la performance humaine et à la cohérence.

Facteur de coûts	Impact commercial
Fluctuations de productivité	Plage de performances de 20 à 50 %
Absentéisme et couverture	15 à 25 % de capacité supplémentaire sont nécessaires
Cycles de fatigue et de motivation	Variation de productivité de 10 à 30 %
Incohérences dans les procédures	Perte d'efficacité de 10 à 40 %
Variations du contrôle qualité	10 à 30 % du coût total
Frais de coordination	15 à 25 % des coûts d'exploitation

Coûts de technologie et d'infrastructure

Les plateformes de gestion des licences indiquent les coûts des logiciels et les taux d'utilisation. La surveillance de l'infrastructure fournit des données de disponibilité, des indicateurs de performance et des coûts de maintenance. Les systèmes de centre d'assistance suivent les frais de support et les problèmes techniques récurrents. Les systèmes de gestion des fournisseurs capturent le coût total des relations technologiques, y compris les dépenses d'intégration et les performances des niveaux de service.

Voici une liste d'exemples de facteurs de coûts liés à la technologie et à l'infrastructure.

Facteur de coûts	Impact commercial
Systèmes technologiques	50 à 500$ par utilisateur et par mois

Espace de travail et équipement	200 à 1 000$ par employé et par mois

Coûts liés aux opportunités commerciales perdues

Les plateformes CRM contiennent des informations sur les temps de réponse aux prospects, les taux de conversion et les opportunités manquées. L'automatisation du marketing montre que les délais de suivi ont un impact sur les conversions de prospects. Les systèmes de support client révèlent comment les problèmes opérationnels affectent la satisfaction et la fidélisation. L'analyse concurrentielle fournit les exigences de réponse du marché et des données sur les gains ou les pertes qui relient les performances opérationnelles aux résultats en termes de revenus.

Voici une liste d'exemples de facteurs de coûts liés à la perte d'opportunités commerciales.

Facteur de coûts	Impact commercial
Retards de réponse du marché	Revenu par jour de retard
Contraintes de capacité	Opportunités commerciales perdues
Allocation de ressources pour l'innovation	Coût d'opportunité du travail de routine
Retards d'acquisition de clients	50 à 90 % de perte de plomb en cas de réponse lente

Coûts des risques et des défauts

La documentation des polices d'assurance indique les coûts de la responsabilité civile générale, de la responsabilité professionnelle, de l'indemnisation des accidents du travail et de la couverture de cyberresponsabilité. Les rapports internes d'évaluation des risques identifient les vulnérabilités opérationnelles et les coûts d'atténuation associés. Les systèmes de suivi des défauts documentent les défaillances des produits ou des services, notamment les coûts de détection, les frais de remplacement et les demandes de garantie. Les calendriers de remplacement des actifs indiquent les taux de défaillance de l'équipement et les coûts de remplacement. Les rapports d'incidents de sécurité font le suivi des accidents du travail et des demandes d'indemnisation des travailleurs connexes. Les plans de continuité des activités détaillent les coûts des systèmes de sauvegarde et les investissements dans la reprise après sinistre.

Voici une liste d'exemples de facteurs de coûts liés aux risques et aux défauts.

Facteur de coûts	Incidence commerciale
Coûts d'assurance	1 à 5 % du budget opérationnel
Coût des erreurs	50 à 5 000$ par incident d'erreur
Impact de l'erreur humaine	2 à 15 % du coût d'exploitation total
Taux d'erreur et retouches	1,5 à 4 fois le coût initial des corrections

Modèles réussis pour la mise en œuvre de systèmes d'IA agentic sur AWS

L'état de l'adoption de l'IA dans les entreprises (rapport ISG 2025) révèle que le principal obstacle à la réussite de la mise en œuvre de l'IA n'est pas la capacité technique, mais le manque d'apprentissage. Ce terme désigne les systèmes qui ne peuvent pas s'adapter, mémoriser le contexte ou s'améliorer au fil du temps. Organisations qui mettent en œuvre des outils d'IA statiques enregistrent des taux d'échec élevés. Voici les caractéristiques communes des systèmes d'intelligence artificielle agentiques qui réussissent :

- Mémoire contextuelle : systèmes qui conservent l'historique des conversations et les préférences de l'utilisateur
- Intégration du feedback — Capacité à tirer des leçons des corrections et à améliorer les performances
- Adaptation du flux de travail — Ajustement automatique à l'évolution des exigences commerciales
- Amélioration continue — Amélioration mesurable grâce à l'expérience opérationnelle

Organisations qui mettent en œuvre l'IA avec succès accordent souvent la priorité aux éléments suivants :

- Utiliser des écosystèmes de partenaires complets plutôt que de créer et d'explorer de manière indépendante les capacités de l'IA
- Des systèmes capables d'apprentissage plutôt que des outils statiques

- Mettre l'accent sur les résultats commerciaux plutôt que sur la comparaison des caractéristiques techniques
- Intégration des flux de travail plutôt que des outils autonomes
- Adaptation continue plutôt qu'une mise en œuvre ponctuelle

Ces modèles s'alignent sur de nombreuses Service AWS fonctionnalités, en particulier l'accès au modèle de base dans Amazon Bedrock, l'architecture basée sur les événements et la surveillance complète proposée par Amazon. AWS Lambda CloudWatch Pour plus d'informations sur l'intégration de la rétroaction humaine et des systèmes capables d'apprentissage, consultez la section Intégration de la rétroaction humaine dans les systèmes d'IA agentique de ce guide.

Intégrer le feedback humain dans les systèmes d'IA agentic

Aucun système ne fonctionne à 100 %, et une défaillance est inévitable. Chaque échec entraîne un coût associé au changement. L'approche Human in the Loop est une approche basée sur l'IA dans laquelle l'IA exécute une tâche, mais l'intervention ou l'approbation d'un être humain est requise. Cette approche doit être utilisée lorsque le coût d'une défaillance est supérieur au coût d'une human-in-the-loop solution.

Le succès des systèmes d'IA agentic dépend fondamentalement de la capacité de l'agent à apprendre et à s'améliorer grâce au feedback humain. Le coût de l'effort humain doit être pris en compte, en fonction du niveau d'effort requis. Contrairement aux outils d'automatisation statiques qui exécutent des règles prédéterminées, les human-in-the-loop solutions sont dotées de systèmes agentiques capables d'apprentissage qui créent un partenariat dynamique entre les agents autonomes et l'humain. L'expertise humaine améliore continuellement les performances de l'agent tandis que les agents gèrent les traitements de routine à grande échelle. Cette approche collaborative transforme la mise en œuvre de l'IA d'un déploiement ponctuel en un processus d'optimisation continu. Le système s'adapte aux modèles organisationnels, internalise les normes de qualité et affine ses capacités de prise de décision sur la base d'une expérience opérationnelle réelle. En capturant systématiquement les corrections, les approbations et les informations humaines, les entreprises peuvent créer des agents d'intelligence artificielle capables de comprendre le contexte, de reconnaître les modèles et de s'aligner de plus en plus sur les objectifs commerciaux au fil du temps.

Pour les solutions qui ne nécessitent pas d'intervention ou de soutien humains, il n'est pas nécessaire de prendre en compte les coûts spécifiques à l'homme dans l'économie des agents.

Apprentissage comportemental par des opérateurs humains

Les opérateurs humains fournissent des informations critiques que les systèmes d'IA agentiques peuvent utiliser pour apprendre, adapter et améliorer leurs réponses au fil du temps. Cette boucle de rétroaction crée un environnement collaboratif dans lequel l'expertise humaine améliore les capacités des agents tandis que les agents s'occupent des traitements de routine.

Grâce à la reconnaissance des modèles de comportement humain, les agents tirent des leçons des modèles d'interaction humaine pour refléter les approches de communication efficaces. Cela les aide à s'adapter aux modèles de décision organisationnels et aux niveaux de tolérance au risque. Les systèmes internalisent les attentes en matière de qualité par le biais de corrections et d'approbations humaines. Ils peuvent également apprendre les réponses appropriées aux différents segments de clientèle et contextes commerciaux.

Les mécanismes de collecte de commentaires efficaces capturent systématiquement les modifications apportées par l'homme aux réponses des agents. Ils analysent ce que les réviseurs humains approuvent, rejettent ou modifient dans les recommandations des agents. En comprenant pourquoi certains cas nécessitent une intervention humaine et en intégrant une évaluation humaine des performances des agents selon différents scénarios et niveaux de complexité, ces systèmes affinent continuellement leurs capacités afin de mieux s'aligner sur les normes et les attentes de l'organisation.

Opérations d'apprentissage continu

L'intégration de l'apprentissage en temps réel permet aux systèmes d'IA agentic d'intégrer le feedback humain et d'améliorer immédiatement les réponses des agents grâce à la mise à jour dynamique des modèles. Ces systèmes utilisent les connaissances humaines pour identifier de nouveaux modèles et des cas extrêmes. Cela améliore leurs capacités de reconnaissance des formes tout en renforçant la mémoire organisationnelle grâce à des expériences d'apprentissage guidées par l'homme. Le perfectionnement continu basé sur les commentaires des opérateurs et les résultats commerciaux favorise une optimisation continue des performances.

La formation guidée par l'homme capture les connaissances d'experts pour améliorer les capacités décisionnelles des agents. Il transfère l'expertise essentielle d'opérateurs expérimentés vers le système d'IA. Grâce à l'apprentissage basé sur des scénarios, les systèmes utilisent des exemples créés par l'homme pour améliorer leur gestion de situations complexes. Ils alignent également les normes de performance des agents sur les attentes en matière de qualité humaine grâce à un étalonnage de la qualité. Cette approche intègre des connaissances humaines sur la culture

organisationnelle et les attentes des clients. Cette adaptation culturelle aide les agents à réagir de manière appropriée dans différents contextes.

Excellence opérationnelle grâce à la collaboration homme-IA

L'optimisation automatisée consciente des risques permet une évaluation continue des conditions de fonctionnement et de la probabilité d'erreur avec une supervision humaine pour les scénarios à haut risque. Cela permet aux systèmes de tirer des leçons des évaluations des risques humains et d'améliorer la prise de décisions futures. Amazon Bedrock donne accès à plusieurs modèles de base dotés de capacités et de profils de coûts différents. Cela permet un routage intelligent qui prend en compte à la fois les profils de coût et de risque tout en intégrant le feedback humain pour optimiser la sélection des modèles. Le réglage des performances équilibre l'efficacité avec la minimisation du taux d'erreur en intégrant le feedback humain sur les normes de qualité et les compromis de performance acceptables. Les décisions automatisées tiennent compte du coût total de possession ajusté au risque. Les opérateurs fournissent des conseils sur la tolérance au risque organisationnel et la pondération des priorités commerciales. Cela vous permet d'optimiser les coûts tout en vous alignant sur les objectifs de l'organisation.

Les systèmes d'apprentissage améliorés par l'homme donnent la priorité à l'apport humain en fonction de l'impact des erreurs et des conséquences commerciales. Cela crée des systèmes d'apprentissage qui comprennent à la fois la précision technique et le contexte commercial grâce à un feedback pondéré en fonction des risques. L'analyse régulière des performances intègre des mesures de risque et une analyse des coûts d'erreur, les informations humaines fournissant un contexte que les systèmes automatisés ne peuvent pas saisir. Le développement des meilleures pratiques met l'accent sur la gestion des risques et la prévention des erreurs en combinant la reconnaissance automatique des formes avec l'expertise et le jugement humains. Le renforcement des capacités organisationnelles par le biais de programmes de formation développe à la fois les compétences humaines pour gérer les systèmes d'IA agentiques et les capacités des agents pour soutenir la prise de décision humaine. Cela garantit une approche globale de la collaboration homme-IA qui renforce les deux composantes du partenariat.

Transformation économique vers une tarification basée sur les résultats pour les systèmes d'IA agentiques sur AWS

Le passage des modèles traditionnels à coûts fixes à une tarification basée sur les résultats représente une transformation fondamentale de la manière dont les organisations structurent leurs opérations économiques et gèrent les risques. Cette transformation ouvre la voie à une

modernisation constante des processus existants tout en finançant la transformation de l'IA agentique. Il permet aux organisations de passer d'opérations statiques gourmandes en ressources à des modèles commerciaux dynamiques axés sur les résultats.

Modèle initial traditionnel

Les ministères fonctionnent souvent comme des centres de coûts dont les coûts directs de main-d'œuvre sont financés par la répartition des coûts. Organisations souhaitent généralement réduire cette allocation de coûts. Si le processus n'est pas modernisé, le ministère doit obtenir les mêmes résultats avec un effectif réduit. Cela dégrade généralement la qualité. Les modèles commerciaux traditionnels créent des défis importants, notamment :

- Évolution linéaire des coûts en fonction de l'augmentation des volumes : cela oblige les entreprises à recruter du personnel supplémentaire pour gérer l'augmentation du volume.
- Engagements en matière de coûts fixes : ils persistent quelles que soient les performances de l'entreprise et l'efficacité des processus.
- Planification avancée — La flexibilité limitée en période de ralentissement économique et de contraintes de capacité nécessite une planification préalable.
- Cycle de dégradation de la qualité — La réduction des budgets entraîne une baisse de la qualité du service lorsque les coûts sont réduits sans amélioration des processus.

Modèle basé sur les résultats

Les modèles modernes basés sur les résultats lient directement les paiements à des résultats commerciaux mesurables, tels que les recrutements réussis, les indicateurs de qualité atteints, les améliorations de l'efficacité des processus ou les gains de productivité réalisés. Cela déplace fondamentalement le risque financier des unités commerciales vers les fournisseurs de services, tout en créant un alignement naturel des incitations. Les principaux avantages d'un modèle basé sur les résultats sont les suivants :

- Les coûts évoluent directement en fonction de la valeur commerciale générée
- Alignement naturel entre les dépenses d'exploitation et les recettes
- Flexibilité pour ajuster la capacité en fonction des conditions du marché
- Pay-per-success les modèles réduisent le risque financier en faisant passer l'exposition financière de l'investissement initial à une performance opérationnelle continue

- Concentrez-vous sur des systèmes capables d'apprentissage qui s'améliorent au fil du temps, plutôt que sur des alternatives statiques

Cette transformation va bien au-delà des centres de coûts internes et redéfinit fondamentalement la manière dont les organisations interagissent avec leurs partenaires et fournisseurs de services externes. En appliquant une tarification basée sur les résultats aux collaborations avec les partenaires, les organisations peuvent améliorer la qualité à long terme et réduire les coûts tout en mettant indirectement l'accent sur la modernisation de l'IA agentique.

Organisations peuvent expérimenter rapidement, mesurer clairement les performances et évoluer en fonction de la valeur commerciale réelle générée plutôt que des engagements traditionnels en ressources fixes. Cette approche permet de :

- Évolution de la relation avec les fournisseurs — Les partenaires s'investissent dans le succès des clients plutôt que dans la simple prestation de services.
- Indicateurs de résultats standardisés — Simplifiez les processus d'approvisionnement auprès de plusieurs fournisseurs.
- Réactivité du marché — Adaptez-vous rapidement à l'évolution des conditions du marché et aux besoins des clients.
- Avantage concurrentiel — Utilisation supérieure des ressources et capacités opérationnelles améliorées.
- Partenariats axés sur la qualité — La collaboration à long terme est axée sur l'amélioration continue et sur des résultats mesurables.

Utilisation AWS Marketplace comme pay-per-outcome activateur

Le principal moteur de cette transformation est AWS Marketplacecelui qui sert de véhicule transactionnel pour le travail d'agence et la tarification basée sur les résultats. Il donne accès à des centaines d'agents d'intelligence artificielle prédéfinis et à des solutions agentiques avec des modèles de tarification transparents basés sur l'utilisation. Cela peut aider à éliminer les coûts de licence initiaux, à réduire la complexité de la mise en œuvre et à permettre aux entreprises de se concentrer sur des systèmes capables d'apprentissage qui s'adaptent et s'améliorent au fil du temps plutôt que sur des alternatives statiques

L'utilisation AWS Marketplace peut apporter les avantages suivants :

- Expérimentation rapide — Testez plusieurs solutions sans investissement en capital important

- Tarification transparente — Coûts basés sur l'utilisation avec une attribution claire aux résultats commerciaux
- Solutions éprouvées — Accès à des agents éprouvés provenant de fournisseurs expérimentés
- Intégration intégrée — Connectivité fluide avec les systèmes existants Services AWS
- Atténuation des risques — Possibilité de changer de fournisseur en fonction des performances
- Accès aux capacités d'apprentissage — Disponibilité de systèmes adaptatifs sans coûts de développement internes

Cette approche permet aux organisations de comparer plusieurs options en fonction de l'obtention des résultats et des capacités d'apprentissage plutôt que de listes de fonctionnalités. Cela peut également vous aider à établir des critères de réussite et des méthodologies de mesure clairs et à négocier une tarification basée sur les résultats, liée aux résultats commerciaux et à l'amélioration du système. En finançant la transformation de l'IA agentique par le biais de modèles basés sur les résultats, les entreprises peuvent moderniser leurs processus en permanence tout en ne payant que pour des améliorations mesurables et des résultats positifs.

Étude de cas : comparaison des coûts de l'IA humaine et agentique pour les opérations de recrutement

Les opérations de recrutement constituent une étude de cas convaincante pour évaluer les compromis économiques entre les systèmes d'IA humains et agentiques, mais le calcul du retour sur investissement dépend essentiellement de votre base opérationnelle actuelle. Organisations évaluant les investissements dans l'IA agentique se posent souvent une question fondamentale : « Et si nous optimisions simplement nos processus humains existants ? » Pour y remédier directement, cette analyse présente deux scénarios distincts qui se situent entre les différentes catégories d'efficacité opérationnelle humaine.

Le scénario A modélise des durées de 45 minutes pour examiner un curriculum vitae (CV) ou un CV. Le scénario B montre des opérations humaines optimisées à 15 minutes par application, ce qui représente une amélioration de 66 % de l'efficacité. Par exemple, cette amélioration peut être obtenue grâce à des processus rationalisés, à des recruteurs expérimentés ou à des outils spécialisés.

En comparant les capacités de systèmes d'agents identiques à ces différents niveaux de référence de performance humaine, nous révélons l'impact de l'efficacité des processus existants sur le calcul du retour sur investissement, les délais d'atteinte du seuil de rentabilité et les décisions de mise en œuvre stratégiques. Cette approche à deux scénarios répond à de multiples objectifs. Cela empêche les organisations de rejeter l'IA agentique en supposant que l'optimisation des processus est suffisante à elle seule. Il aide également les organisations dont les processus sont déjà efficaces à comprendre leurs spécificités économiques. En outre, ces scénarios mettent en évidence les cas dans lesquels les avantages non financiers, tels que la disponibilité 24 heures sur 24 et 7 jours sur 7 et l'évolutivité, deviennent les principaux facteurs de décision. La compréhension de ces dynamiques économiques selon différents critères d'efficacité permet aux entreprises de prendre des décisions éclairées quant au lieu et au moment de déployer des systèmes d'IA agentic pour un impact commercial maximal.

Scénario A : 45 minutes de projection

Le scénario A représente des opérations de recrutement au cours desquelles les recruteurs humains passent 45 minutes à examiner chaque CV. Ce scénario modélise un recruteur de niveau intermédiaire dont le coût annuel complet est de 112 250$. Ce recruteur traite les candidatures pendant les heures normales de bureau avec des caractéristiques de performance humaine typiques.

En revanche, le système d'intelligence artificielle agentic nécessite un investissement initial de 23 000 dollars pour le développement, la personnalisation et l'intégration de l'ATS, et ses coûts d'exploitation mensuels minimaux sont de 500 dollars pour l'infrastructure cloud. L'agent traite les demandes en seulement 5 minutes avec une disponibilité 24 heures sur 24, 7 jours sur 7, avec un taux d'erreur de 2 % et une capacité mensuelle supérieure à 8 600 applications. Il s'agit d'un écart d'efficacité considérable, car l'agent fonctionne 9 fois plus vite par application et sa capacité mensuelle est 39 fois supérieure. Cette section examine l'analyse de la structure des coûts, les indicateurs opérationnels, les comparaisons basées sur les volumes et les calculs du retour sur investissement cumulé au cours des six premiers mois d'exploitation.

Structure des coûts de base

Le tableau suivant indique les coûts de configuration initiaux pour le scénario A.

Composant	Opérations humaines	Système Agentic AI
Développement et personnalisation des agents	N/A	15 000$
Intégration du système de suivi des candidats (ATS)	N/A	5 000$
Formation et optimisation	N/A	3 000$
Coût d'installation initial total	0$	23 000$

Le tableau suivant indique les coûts fixes annuels pour le scénario A.

Composant	Opérations humaines	Système Agentic AI
Salaire de base	65 000$	N/A
Avantages (30 %)	19 500$	N/A
Espace de travail et équipement	12 000$	N/A

Supervision de la gestion (15 %)	9 750$	N/A
Formation et développement	6 000$	N/A
Coût fixe annuel total	112 250$	S/O

Le tableau suivant indique les coûts d'exploitation mensuels pour le scénario A.

Composant	Opérations humaines	Système Agentic AI
Informatique en nuage	N/A	$200
Stockage	N/A	100 USD
Opérations de base de données	N/A	100 USD
Contrôle	N/A	100 USD
Coût fixe mensuel total	9 354$	500$

Métriques opérationnelles

Le tableau suivant présente les mesures opérationnelles pour le scénario A.

Métrique	Opérations humaines	Système Agentic AI
Délai de traitement par demande	45 minutes	5 minutes
Capacité horaire	1.33 demandes	12 demandes
Capacité journalière (24 heures)	10 à 11 demandes	288 demandes
Capacité mensuelle	220 demandes	8 640 demandes

Coût par demande	45$	2,50$
Coût par embauche réussie	2 200$	$125
Taux d'erreur	5 %	2 %
Coût de correction d'erreur	90$ par erreur	45$ par escalade

Analyse des coûts basée sur le volume

Le tableau suivant présente une analyse des coûts basée sur le volume pour le scénario A. Dans cet exemple, le coût du système d'IA agentic inclut les coûts fixes et les coûts d'installation amortis de 1 917$ par mois sur 12 mois.

Volume mensuel	Coût humain	Coût du système Agentic AI	Économies mensuelles
100 demandes	4 500$	750$	3 750$
500 demandes	22 500$	2 667$	19 833$
1 000 demandes	45 000$	4 917$	40 083$

Analyse du retour sur investissement

Le tableau suivant présente une analyse du retour sur investissement pour le scénario A basée sur le traitement de 500 demandes par mois.

Métrique	Valeur
Coût humain mensuel	22 500$
Coût mensuel de l'agent	2 667$
Économies mensuelles	19 833$
Économies annuelles	237 996$

Période d'équilibre	1,16 mois

Comparaison des coûts cumulés

Le tableau suivant présente une comparaison des coûts cumulés pour le scénario A pour les six premiers mois, en supposant 500 demandes par mois.

Mois	Coût humain	Coût du système Agentic AI	Économies cumulées
1	22 500$	25 667$	-3 167$
2	45 000$	28 334$	16 666$
3	67 500$	31 001$	36 499$
4	90 000$	33 668$	56 332$
5	112 500$	36 335$	76 165$
6	135 000$	39 002$	95 998$

Avantages supplémentaires du système d'IA agentic

Les avantages supplémentaires fournis par le système d'IA agentic dans le scénario A sont les suivants :

- Évolutivité — Peut gérer les pics de volume sans frais supplémentaires
- Disponibilité : fonctionnement 24 heures sur 24, 7 jours sur 7 avec réponse immédiate
- Cohérence — Application uniforme des critères de sélection
- Efficacité du temps — Réduction significative time-to-hire
- Expérience utilisateur — Feedback instantané aux candidats

Scénario B : 15 minutes de projection

Les modèles du scénario B ont optimisé les opérations de recrutement dans le cadre desquelles les recruteurs humains ont rationalisé leur processus de sélection à 15 minutes par candidature. Cela représente une amélioration de 66 % de l'efficacité par rapport au scénario A. Ce scénario maintient le même coût annuel complet de 112 250$ pour un recruteur de niveau intermédiaire. Cependant, il démontre une amélioration significative de la productivité humaine, la capacité quotidienne passant à 32 applications sur une période de 8 heures et le débit mensuel atteignant 660 applications. L'amélioration de l'efficacité humaine réduit le coût par application de 45$ à 15$, réduisant ainsi l'écart économique avec le système d'IA agentic. L'agent conserve toutefois ses avantages structurels : temps de traitement de 5 minutes, disponibilité 24 heures sur 24, 7 jours sur 7, permettant d'exécuter 288 applications quotidiennes, taux d'erreur inférieur de 2 % par rapport aux 5 % humains et capacité mensuelle supérieure à 8 600 applications. Bien que cette amélioration de l'efficacité prolonge le seuil de rentabilité de 1,16 mois à 4,76 mois et réduise les économies mensuelles de 19 833 dollars à 4 833 dollars, l'analyse révèle que les systèmes d'agents restent économiquement viables même lorsqu'ils sont en concurrence avec des opérations humaines hautement optimisées. Il s'agit là d'une information essentielle pour les entreprises qui évaluent si les niveaux actuels d'efficacité de leurs processus justifient un investissement dans l'IA agentique.

Structure des coûts de base

Le tableau suivant indique les coûts fixes annuels pour le scénario B.

Composant	Opérations humaines	Système Agentic AI
Salaire de base	65 000$	N/A
Avantages (30 %)	19 500$	N/A
Espace de travail et équipement	12 000$	N/A
Supervision de la gestion (15 %)	9 750$	N/A
Formation et développement	6 000$	N/A
Coût fixe annuel total	112 250$	S/O

Le tableau suivant indique les coûts de mise en œuvre du scénario B.

Composant	Opérations humaines	Système Agentic AI
Configuration initiale du	N/A	23 000$
Coûts fixes mensuels	9 354$	500$

Métriques opérationnelles

Le tableau suivant présente les mesures opérationnelles pour le scénario B.

Métrique	Opérations humaines	Système Agentic AI
Délai de traitement par demande	15 minutes	5 minutes
Capacité horaire	4 demandes	12 demandes
Capacité quotidienne (équipe de 8 heures)	32 demandes	288 demandes
Capacité mensuelle	660 demandes	8 640 demandes
Coût par demande	15$	2,50$
Coût par embauche réussie	2 200$	$125
Taux d'erreur	5 %	2 %
Coût de correction d'erreur	30$ par erreur	45$ par escalade

Analyse des coûts basée sur le volume

Le tableau suivant présente une analyse des coûts basée sur le volume pour le scénario B. Dans cet exemple, le coût du système d'IA agentic inclut les coûts fixes et les coûts d'installation amortis de 1 917$ par mois sur 12 mois.

Volume mensuel	Coût humain	Coût du système Agentic AI	Économies mensuelles
100 demandes	1 500$	750$	750$
500 demandes	7 500$	2 667$	4 833$
1 000 demandes	15 000$	4 917$	10 083$

Analyse du retour sur investissement

Le tableau suivant présente une analyse du retour sur investissement pour le scénario B basée sur le traitement de 500 demandes par mois.

Métrique	Valeur
Coût humain mensuel	7 500$
Coût mensuel du système d'IA agentic	2 667$
Économies mensuelles	4 833$
Économies annuelles	57 996$
Période d'équilibre	4,76 mois

Comparaison des coûts cumulés

Le tableau suivant présente une comparaison des coûts cumulés pour le scénario B pour les six premiers mois, en supposant 500 demandes par mois.

Mois	Coût humain	Coût du système Agentic AI	Économies cumulées
1	7 500$	25 667$	-18 167$
2	15 000$	28 334$	-13 334$

3	22 500$	31 001$	-8 501 dollars
4	30 000$	33 668$	-3 668$
5	37 500 dollars	36 335$	1 165$
6	45 000$	39 002$	5 998$

Comparaison des coûts et des avantages pour chaque scénario

Métrique	Scénario A	Scénario B	Impact
Heure de projection	45 minutes	15 minutes	Amélioration de 66 %
Capacité quotidienne	10 à 11 demandes	32 demandes	Augmentation de 200 %
Coût par demande	45$	15$	66 % de réduction
Économies mensuelles (500 demandes)	19 833$	4 833$	Diminution de 76 %
Période d'équilibre	1,16 mois	4,76 mois	310 % plus long

Le scénario B montre des gains d'efficacité significatifs dans les opérations humaines, avec des améliorations du temps de traitement qui augmentent la capacité sans personnel supplémentaire et réduisent considérablement le coût par application. Cependant, l'impact financier donne une image plus nuancée : si le retour sur investissement reste positif, les entreprises sont confrontées à une période d'équilibre prolongée et à des économies mensuelles réduites par rapport au scénario A. Ces résultats mettent en évidence des facteurs décisionnels essentiels pour la mise en œuvre : le système d'agents reste financièrement viable même avec des opérations humaines optimisées, mais les organisations doivent adopter une perspective d'investissement à plus long terme et prendre soigneusement en compte les fluctuations de volume et les besoins d'évolutivité lors de l'évaluation des délais de déploiement et des rendements attendus.

Cependant, le système d'intelligence artificielle agentic conserve des avantages opérationnels essentiels qui vont au-delà des simples économies de coûts. Il assure une disponibilité 24 heures

sur 24, 7 jours sur 7 pour un engagement immédiat des candidats, quels que soient les fuseaux horaires ou les heures de bureau. Il fournit une qualité de criblage constante en appliquant des critères uniformes à chaque application, des échelles permettant de gérer les pics de volume sans encourir de coûts supplémentaires. Il offre une réponse immédiate aux candidats qui améliore la marque employeur et l'expérience des candidats, et fonctionne avec un facteur de fatigue nul qui garantit des performances de haute qualité à la première candidature comme à la millième.

Les erreurs humaines sont généralement le résultat de la fatigue, de la distraction ou de lacunes dans les connaissances et impliquent souvent une mauvaise communication ou des informations incorrectes. Les erreurs du système Agentic AI proviennent généralement de cas extrêmes, de saisies ambiguës ou de limites des données d'entraînement. Ces erreurs ont tendance à être de nature plus constante.

Les indicateurs de qualité et d'expérience révèlent des compromis évidents entre les capacités humaines et celles des agents :

- Satisfaction du client — Les humains excellent dans l'empathie et la résolution de problèmes complexes, et les agents fournissent des informations cohérentes et précises pour les requêtes de routine.
- Temps de réponse — Le temps de réponse favorise les agents avec une disponibilité immédiate 24 heures sur 24, 7 jours sur 7. Des humains fournissent une assistance pendant les heures ouvrables, ce qui peut entraîner des retards dans les files d'attente.
- Cohérence : les agents fournissent des réponses identiques à des requêtes similaires. Les humains peuvent avoir des approches et des applications des connaissances différentes.
- Gestion de l'escalade — Les problèmes complexes qui nécessitent du jugement, de la créativité ou de l'intelligence émotionnelle restent les points forts de l'être humain.

Conclusion et ressources

L'économie des systèmes humains par rapport aux systèmes d'IA agentiques représente bien plus qu'une simple décision technologique. Il reflète une transformation fondamentale de la manière dont les organisations créent de la valeur, gèrent les risques et obtiennent un avantage concurrentiel. Le succès nécessite une évaluation systématique des caractéristiques des emplois, une mesure complète des résultats (y compris les facteurs de risque) et une mise à l'échelle stratégique basée sur des résultats prouvés.

L'état de l'adoption de l'IA dans les entreprises (rapport ISG 2025) révèle que la plupart des implémentations de l'IA échouent en raison de lacunes d'apprentissage, c'est-à-dire de systèmes incapables de s'adapter, de mémoriser le contexte ou de s'améliorer au fil du temps. Organisations qui réussissent se concentrent sur des systèmes capables d'apprentissage qui s'intègrent parfaitement aux flux de travail et démontrent une amélioration continue grâce au feedback humain et à l'expérience opérationnelle.

Organisations qui comprennent ces principes (en commençant par les emplois appropriés, en décomposant les emplois en tâches, en mesurant tout, y compris l'impact sur les risques, et en adaptant ce qui fonctionne) obtiendront un avantage concurrentiel durable grâce à une utilisation optimale des ressources et à une automatisation axée sur les résultats qui évolue avec le succès de l'entreprise.

L'avenir appartient aux organisations capables de combiner intelligemment l'expertise humaine avec les capacités d'intelligence artificielle agentique. Cela crée des modèles hybrides qui fournissent des résultats supérieurs tout en maintenant la flexibilité, la capacité d'apprentissage et les avantages collaboratifs nécessaires aux conditions dynamiques du marché.

Ressources

Les ressources suivantes peuvent vous aider à planifier, concevoir et mettre en œuvre des systèmes d'IA agentic sur : AWS

- Création d'architectures sans serveur pour l'IA agentic (directives prescriptives AWS)AWS
- Opérationnalisation de l'IA agentique sur AWS(directives prescriptives)AWS
- Modèles et flux de travail liés à l'IA agentic AWS(directivesAWS prescriptives)
- Agentic AI (directivesAWS prescriptives)

- Hub d'optimisation des coûts AWS (Service AWS)
- Documentation Amazon Bedrock ()Service AWS
- Pilier d'optimisation des coûts (AWS Well-Architected Framework)
- Agents et solutions d'IA (AWS Marketplace)

Historique du document

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

Modification	Description	Date
Publication initiale	—	28 janvier 2026

AWS Glossaire des directives prescriptives

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

Nombres

7 R

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

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

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

A

ABAC

Voir contrôle d'accès basé sur les attributs.

services abstraits

Consultez la section Services gérés.

ACIDE

Voir atomicité, consistance, isolation, durabilité.

migration active-active

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

migration active-passive

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

fonction d'agrégation

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

AI

Voir intelligence artificielle.

AIOps

Voir les opérations d'intelligence artificielle.

anonymisation

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

anti-motif

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

contrôle des applications

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

portefeuille d'applications

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

intelligence artificielle (IA)

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

opérations d'intelligence artificielle (AIOps)

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

chiffrement asymétrique

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

atomicité, cohérence, isolement, durabilité (ACID)

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

contrôle d'accès par attributs (ABAC)

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

source de données faisant autorité

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

Zone de disponibilité

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

AWS Cadre d'adoption du cloud (AWS CAF)

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

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

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

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

B

mauvais bot

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

BCP

Consultez la section Planification de la continuité des activités.

graphique de comportement

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

système de poids fort

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

classification binaire

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

filtre de Bloom

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

déploiement bleu/vert

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

bot

Application logicielle qui exécute des tâches automatisées sur Internet et simule l'activité ou l'interaction humaine. Certains robots sont utiles ou bénéfiques, comme les robots d'exploration Web qui indexent des informations sur Internet. D'autres robots, appelés « bots malveillants », sont destinés à perturber ou à nuire à des individus ou à des organisations.

botnet

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

branche

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

accès par brise-vitre

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

stratégie existante (brownfield)

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

cache de tampon

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

capacité métier

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

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

planification de la continuité des activités (BCP)

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

C

CAF

Voir le cadre d'adoption du AWS cloud.

déploiement de Canary

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

CCo E

Voir le Centre d'excellence du cloud.

CDC

Voir capture des données de modification.

capture des données de modification (CDC)

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

ingénierie du chaos

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

CI/CD

Découvrez l'intégration continue et la livraison continue.

classification

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

chiffrement côté client

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

Centre d'excellence du cloud (CCoE)

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

cloud computing

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

modèle d'exploitation du cloud

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

étapes d'adoption du cloud

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

- Projet : exécution de quelques projets liés au cloud à des fins de preuve de concept et d'apprentissage
- Base : réaliser des investissements fondamentaux pour accélérer votre adoption du cloud (par exemple, créer une zone de landing zone, définir un CCo E, établir un modèle opérationnel)
- Migration : migration d'applications individuelles
- Réinvention : optimisation des produits et services et innovation dans le cloud

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

CMDB

Consultez la base de données de gestion des configurations.

référentiel de code

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

cache passif

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

données gelées

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

vision par ordinateur (CV)

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

dérive de configuration

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

base de données de gestion des configurations (CMDB)

Référentiel qui stocke et gère les informations relatives à une base de données et à son environnement informatique, y compris les composants matériels et logiciels ainsi que leurs

configurations. Vous utilisez généralement les données d'une CMDB lors de la phase de découverte et d'analyse du portefeuille de la migration.

pack de conformité

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

intégration continue et livraison continue (CI/CD)

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

CV

Voir vision par ordinateur.

D

données au repos

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

classification des données

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

dérive des données

Une variation significative entre les données de production et les données utilisées pour entraîner un modèle ML, ou une modification significative des données d'entrée au fil du temps. La dérive

des données peut réduire la qualité, la précision et l'équité globales des prédictions des modèles ML.

données en transit

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

maillage de données

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

minimisation des données

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

périmètre de données

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

prétraitement des données

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

provenance des données

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

sujet des données

Personne dont les données sont collectées et traitées.

entrepôt des données

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

langage de définition de base de données (DDL)

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

langage de manipulation de base de données (DML)

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

DDL

Voir langage de définition de base de données.

ensemble profond

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

deep learning

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

defense-in-depth

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

administrateur délégué

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

déploiement

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

environnement de développement

Voir environnement.

contrôle de détection

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

cartographie de la chaîne de valeur du développement (DVSM)

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

jumeau numérique

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

tableau des dimensions

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

catastrophe

Un événement qui empêche une charge de travail ou un système d'atteindre ses objectifs commerciaux sur son site de déploiement principal. Ces événements peuvent être des

catastrophes naturelles, des défaillances techniques ou le résultat d'actions humaines, telles qu'une mauvaise configuration involontaire ou une attaque de logiciel malveillant.

reprise après sinistre (DR)

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

DML

Voir langage de manipulation de base de données.

conception axée sur le domaine

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

DR

Voir reprise après sinistre.

détection de dérive

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

DVSM

Voir la cartographie de la chaîne de valeur du développement.

E

EDA

Voir analyse exploratoire des données.

EDI

Voir échange de données informatisé.

informatique de périphérie

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

échange de données informatisé (EDI)

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

chiffrement

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

clé de chiffrement

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

endianisme

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

point de terminaison

Voir point de terminaison de service.

service de point de terminaison

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

planification des ressources d'entreprise (ERP)

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

chiffrement d'enveloppe

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

environment

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

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

épopée

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

ERP

Voir Planification des ressources d'entreprise.

analyse exploratoire des données (EDA)

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

F

tableau des faits

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

échouer rapidement

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

limite d'isolation des défauts

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

branche de fonctionnalités

Voir succursale.

fonctionnalités

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

importance des fonctionnalités

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

transformation de fonctionnalité

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

invitation en quelques coups

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

FGAC

Découvrez le contrôle d'accès détaillé.

contrôle d'accès détaillé (FGAC)

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

migration instantanée (flash-cut)

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

FM

Voir le modèle de fondation.

modèle de fondation (FM)

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

G

IA générative

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

blocage géographique

Voir les restrictions géographiques.

restrictions géographiques (blocage géographique)

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

Flux de travail Gitflow

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

image dorée

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

stratégie inédite

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

barrière de protection

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

H

HA

Découvrez la haute disponibilité.

migration de base de données hétérogène

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

haute disponibilité (HA)

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

modernisation des historiens

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

données de rétention

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

migration de base de données homogène

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

données chaudes

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

correctif

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

période de soins intensifs

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

I

IaC

Considérez l'infrastructure comme un code.

politique basée sur l’identité

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

application inactive

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

IIo T

Voir Internet industriel des objets.

infrastructure immuable

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

VPC entrant (d'entrée)

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

migration incrémentielle

Stratégie de basculement dans le cadre de laquelle vous migrez votre application par petites parties au lieu d'effectuer un basculement complet unique. Par exemple, il se peut que vous ne transfériez que quelques microservices ou utilisateurs vers le nouveau système dans un premier temps. Après avoir vérifié que tout fonctionne correctement, vous pouvez transférer

progressivement des microservices ou des utilisateurs supplémentaires jusqu'à ce que vous puissiez mettre hors service votre système hérité. Cette stratégie réduit les risques associés aux migrations de grande ampleur.

Industry 4.0

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

infrastructure

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

infrastructure en tant que code (IaC)

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

Internet industriel des objets (IIoT)

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

VPC d'inspection

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

Internet des objets (IoT)

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

interprétabilité

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

IoT

Voir Internet des objets.

Bibliothèque d'informations informatiques (ITIL)

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

gestion des services informatiques (ITSM)

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

ITIL

Consultez la bibliothèque d'informations informatiques.

ITSM

Voir Gestion des services informatiques.

L

contrôle d'accès basé sur des étiquettes (LBAC)

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

zone de destination

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

grand modèle de langage (LLM)

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

migration de grande envergure

Migration de 300 serveurs ou plus.

LBAC

Voir contrôle d'accès basé sur des étiquettes.

principe de moindre privilège

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

lift and shift

Voir 7 Rs.

système de poids faible

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

LLM

Voir le grand modèle de langage.

environnements inférieurs

Voir environnement.

M

machine learning (ML)

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

branche principale

Voir succursale.

malware

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

services gérés

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

système d'exécution de la fabrication (MES)

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

MAP

Voir Migration Acceleration Program.

mécanisme

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

compte membre

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

MAILLES

Voir le système d'exécution de la fabrication.

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

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

microservice

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

architecture de microservices

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

Programme d'accélération des migrations (MAP)

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

migration à grande échelle

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

usine de migration

Équipes interfonctionnelles qui rationalisent la migration des charges de travail grâce à des approches automatisées et agiles. Les équipes de Migration Factory comprennent généralement des responsables des opérations, des analystes commerciaux et des propriétaires, des ingénieurs de migration, des développeurs et DevOps des professionnels travaillant dans le cadre de sprints.

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

métadonnées de migration

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

modèle de migration

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

Évaluation du portefeuille de migration (MPA)

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

Évaluation de la préparation à la migration (MRA)

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

stratégie de migration

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

ML

Voir apprentissage automatique.

modernisation

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

évaluation de la préparation à la modernisation

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

applications monolithiques (monolithes)

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

MPA

Voir Évaluation du portefeuille de migration.

MQTT

Voir Message Queuing Telemetry Transport.

classification multi-classes

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

infrastructure mutable

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

O

OAC

Voir Contrôle d'accès à l'origine.

OAI

Voir l'identité d'accès à l'origine.

OCM

Voir gestion du changement organisationnel.

migration hors ligne

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

OI

Consultez la section Intégration des opérations.

OLA

Voir l'accord au niveau opérationnel.

migration en ligne

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

OPC-UA

Voir Open Process Communications - Architecture unifiée.

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

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

accord au niveau opérationnel (OLA)

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

examen de l'état de préparation opérationnelle (ORR)

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

technologie opérationnelle (OT)

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

intégration des opérations (OI)

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

journal de suivi d'organisation

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

gestion du changement organisationnel (OCM)

Cadre pour gérer les transformations métier majeures et perturbatrices du point de vue des personnes, de la culture et du leadership. L'OCM aide les organisations à se préparer et à effectuer la transition vers de nouveaux systèmes et de nouvelles politiques en accélérant

l'adoption des changements, en abordant les problèmes de transition et en favorisant des changements culturels et organisationnels. Dans la stratégie de AWS migration, ce cadre est appelé accélération du personnel, en raison de la rapidité du changement requise dans les projets d'adoption du cloud. Pour plus d'informations, veuillez consulter le guide OCM.

contrôle d'accès d'origine (OAC)

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

identité d'accès d'origine (OAI)

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

ORR

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

DE

Voir technologie opérationnelle.

VPC sortant (de sortie)

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

P

limite des autorisations

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

informations personnelles identifiables (PII)

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

PII

Voir les informations personnelles identifiables.

manuel stratégique

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

PLC

Voir contrôleur logique programmable.

PLM

Consultez la section Gestion du cycle de vie des produits.

policy

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

persistance polyglotte

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

évaluation du portefeuille

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

predicate

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

prédicat pushdown

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

contrôle préventif

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

principal

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

confidentialité dès la conception

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

zones hébergées privées

Conteneur contenant des informations sur la manière dont vous souhaitez qu'Amazon Route 53 réponde aux requêtes DNS pour un domaine et ses sous-domaines au sein d'un ou de plusieurs VPCs domaines. Pour plus d'informations, veuillez consulter Working with private hosted zones dans la documentation Route 53.

contrôle proactif

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

gestion du cycle de vie des produits (PLM)

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

environnement de production

Voir environnement.

contrôleur logique programmable (PLC)

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

chaînage rapide

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

pseudonymisation

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

publish/subscribe (pub/sub)

Modèle qui permet 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. Chacun Région AWS est isolé et indépendant des autres pour garantir tolérance aux pannes, stabilité et résilience. Pour plus d'informations, voir Spécifier ce que Régions AWS votre compte peut utiliser.

régression

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

réhéberger

Voir 7 Rs.

version

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

déplacer

Voir 7 Rs.

replateforme

Voir 7 Rs.

rachat

Voir 7 Rs.

résilience

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

politique basée sur les ressources

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

matrice responsable, redevable, consulté et informé (RACI)

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

contrôle réactif

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

retain

Voir 7 Rs.

se retirer

Voir 7 Rs.

Génération augmentée de récupération (RAG)

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

rotation

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

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

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

RPO

Voir l'objectif du point de récupération.

RTO

Voir l'objectif en matière de temps de rétablissement.

runbook

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

S

SAML 2.0

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

SCADA

Voir Contrôle de supervision et acquisition de données.

SCP

Voir la politique de contrôle des services.

secret

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

sécurité dès la conception

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

contrôle de sécurité

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

renforcement de la sécurité

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

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

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

automatisation des réponses de sécurité

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

chiffrement côté serveur

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

Politique de contrôle des services (SCP)

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

point de terminaison du service

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

contrat de niveau de service (SLA)

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

indicateur de niveau de service (SLI)

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

objectif de niveau de service (SLO)

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

modèle de responsabilité partagée

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

SIEM

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

point de défaillance unique (SPOF)

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

SLA

Voir le contrat de niveau de service.

SLI

Voir l'indicateur de niveau de service.

SLO

Voir l'objectif de niveau de service.

split-and-seed modèle

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

SPOF

Voir point de défaillance unique.

schéma en étoile

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

modèle de figuier étrangleur

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

sous-réseau

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

contrôle de supervision et acquisition de données (SCADA)

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

chiffrement symétrique

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

tests synthétiques

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

invite du système

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

T

tags

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

variable cible

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

liste de tâches

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

environnement de test

Voir environnement.

entraînement

Pour fournir des données à partir desquelles votre modèle de ML peut apprendre. Les données d'entraînement doivent contenir la bonne réponse. L'algorithme d'apprentissage identifie des modèles dans les données d'entraînement, qui mettent en correspondance les attributs des données d'entrée avec la cible (la réponse que vous souhaitez prédire). Il fournit un modèle de ML

qui capture ces modèles. Vous pouvez alors utiliser le modèle de ML pour obtenir des prédictions sur de nouvelles données pour lesquelles vous ne connaissez pas la cible.

passerelle de transit

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

flux de travail basé sur jonction

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

accès sécurisé

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

réglage

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

équipe de deux pizzas

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

U

incertitude

Un concept qui fait référence à des informations imprécises, incomplètes ou inconnues susceptibles de compromettre la fiabilité des modèles de ML prédictifs. Il existe deux types

d'incertitude : l'incertitude épistémique est causée par des données limitées et incomplètes, alors que l'incertitude aléatoire est causée par le bruit et le caractère aléatoire inhérents aux données.

tâches indifférenciées

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

environnements supérieurs

Voir environnement.

V

mise à vide

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

contrôle de version

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

Appairage de VPC

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

vulnérabilités

Défaut logiciel ou matériel qui compromet la sécurité du système.

W

cache actif

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

données chaudes

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

fonction de fenêtre

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

charge de travail

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

flux de travail

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

VER

Voir écrire une fois, lire plusieurs.

WQF

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

écrire une fois, lire plusieurs (WORM)

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

Z

exploit Zero-Day

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

vulnérabilité « jour zéro »

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

invite Zero-Shot

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

application zombie

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

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