Évaluation de la charge de travail générative de

AWS Conseils prescriptifs

AWS Conseils prescriptifs: Évaluation de la charge de travail générative de

Table of Contents

Évaluation de la charge de travail générative de

Tabby Ward et Deepak Dixit, Amazon Web Services ()AWS

Novembre 2024 (historique du document)

L'évaluation de la charge de travail d'IA générative est une méthode stratégique visant à évaluer et à améliorer le niveau de préparation d'une organisation à créer ou à mettre à jour ses charges de travail d'IA générative. Cette évaluation est importante car l'intégration de l'IA générative dans les opérations commerciales peut modifier considérablement la façon dont les choses fonctionnent et peut apporter de nouvelles performances et capacités. Cependant, pour réussir à adopter l'IA générative, il est essentiel de bien comprendre les systèmes actuels et d'avoir un plan clair pour le futur.

Les charges de travail génératives liées à l'IA font référence à des tâches informatiques impliquant l'utilisation de modèles d'intelligence artificielle capables de créer de nouveaux contenus, tels que du texte, des images, du code ou d'autres types de données. Ces charges de travail nécessitent généralement une puissance de calcul importante, du matériel spécialisé tel que GPUs, et de grands ensembles de données pour la formation et l'inférence. L'intégration des charges de travail génératives liées à l'IA dans les opérations présente plusieurs défis :

- Exigences en matière d'infrastructure : provisionnement des ressources informatiques importantes et du matériel spécialisé requis par les modèles d'IA générative.
- Gestion des données : garantir la qualité, la confidentialité et la conformité des données lors de la gestion de grands ensembles de données.
- Manque de compétences : manque d'expertise dans les technologies d'intelligence artificielle et le déploiement de modèles.
- Considérations éthiques : remédier aux préjugés, à l'équité et à la transparence dans le contenu généré par l'IA.
- Complexité de l'intégration : intégration fluide de l'IA générative dans les flux de travail existants et les systèmes existants.
- Gestion des coûts : équilibre entre les avantages potentiels et les coûts élevés de mise en œuvre et d'exploitation.

Pour surmonter ces défis, il faut une planification minutieuse, des investissements dans les infrastructures et les talents, ainsi qu'une approche stratégique de la mise en œuvre.

Objectif de ce guide

L'IA générative devient rapidement un élément essentiel dans de nombreux secteurs. Elle offre des opportunités de transformation mais pose également des défis en termes d'intégration, de conformité et d'évolutivité. De nombreuses organisations ont du mal à tirer pleinement parti de l'IA en raison de la faiblesse des bases technologiques, de la résistance au changement et des problèmes de qualité des données. L'évaluation de la charge de travail de l'IA générative permet de relever ces défis en identifiant les exigences de modernisation, en définissant le champ de mise en œuvre et en remettant en question les systèmes et les modes de pensée existants. Il aide également à déterminer le minimum de produits viables (MVPs) et vous aide à développer une architecture de solution cible, garantissant ainsi une approche structurée et stratégique de l'adoption de l'IA.

Ce guide constitue une approche structurée pour aider les organisations à faire face aux complexités liées à l'adoption de technologies d'IA générative. Au lieu de définir clairement les exigences dès le départ, le guide aide à :

- Identifier les cas d'utilisation potentiels de l'IA générative au sein de votre organisation.
- Évaluer l'état de préparation de votre entreprise à l'adoption de l'IA générative.
- Définition et affinement des objectifs des cas d'utilisation et des objectifs ambitieux.
- Déterminer la portée et les exigences de la mise en œuvre de l'IA générative.
- Développement d'une architecture de solution cible.

Public cible et avantages

Cette évaluation est spécialement conçue pour les architectes de solutions, les architectes d'entreprise et les architectes d'applications qui souhaitent évaluer les aspects techniques de la modernisation de la charge de travail générative de l'IA. Il est également utile pour les responsables des programmes et des personnes qui souhaitent évaluer l'état de préparation global de leur équipe, l'allocation des ressources et les exigences en matière d'habilitation. Les meilleures pratiques du secteur soulignent l'importance d'une évaluation complète pour garantir l'état de préparation à l'adoption de l'IA. Cela inclut l'évaluation de l'architecture, du stockage, de la conformité, de l'intégration, des tests, du déploiement et de l'automatisation.

Portée

Les sujets suivants sont concernés par la méthode d'évaluation de la charge de travail de l'IA générative :

- Technologies et modèles actuels d'IA générative (par exemple, grands modèles de langage, modèles de génération d'images)
- Applications d'IA restreintes qui utilisent des techniques génératives
- Intégration de l'IA générative aux systèmes et flux de travail existants
- Stratégies de données pour la formation et le réglage précis des modèles d'IA générative
- Considérations éthiques et pratiques responsables en matière d'IA pour les applications d'IA générative actuelles
- Stratégies de test et de déploiement pour l'IA générative dans les environnements de production
- Considérations relatives à la sécurité et à la confidentialité pour les implémentations d'IA générative
- Optimisation des performances et évolutivité des charges de travail génératives liées à l'IA
- Cas d'utilisation et applications de l'IA générative dans divers secteurs
- Évaluation des résultats de l'IA générative et des processus d'assurance qualité

Les sujets suivants sont hors de portée :

- Scénarios d'intelligence générale artificielle (AGI) et de superintelligence artificielle (ASI)
- Les futures avancées spéculatives de l'IA au-delà des modèles génératifs actuels
- Applications de l'informatique quantique dans le domaine de l'IA
- Informatique neuromorphique et interfaces cerveau-ordinateur
- Conscience et conscience de soi dans les systèmes d'IA
- Impacts sociétaux à long terme de l'IA avancée au-delà des applications d'IA générative actuelles
- Cadres réglementaires pour les futures technologies d'IA hypothétiques
- Débats philosophiques sur la nature de l'intelligence et de la conscience dans les machines
- Cas extrêmes ou cas d'utilisation hautement spéculatifs de l'IA
- Spécifications techniques détaillées des modèles ou architectures d'IA propriétaires

Résultats commerciaux ciblés

L'évaluation de la charge de travail de l'IA générative vise à obtenir plusieurs résultats ciblés qui sont essentiels pour moderniser avec succès les charges de travail de l'IA générative. Ces résultats garantissent que les organisations sont bien préparées à intégrer les technologies d'IA de manière efficace et efficiente.

Pour chaque résultat visé, l'évaluation de la charge de travail de l'IA générative se concentre sur :

- Interdépendances : Identifiez et clarifiez les interdépendances entre le résultat et les autres aspects du processus de modernisation. Cela implique de comprendre comment un résultat peut influencer ou être influencé par d'autres, afin de garantir une approche holistique de la modernisation.
- Harmonisation des parties prenantes : Décrivez des stratégies pour aligner les différentes parties prenantes sur chaque résultat. Cela implique de communiquer la valeur et l'impact de chaque résultat aux différents niveaux organisationnels et départements, afin de favoriser l'adhésion et le soutien.
- Hiérarchisation : dans les cas où plusieurs cas d'utilisation ou résultats sont identifiés, fournissez un cadre pour les hiérarchiser en fonction de facteurs tels que l'impact commercial, les besoins en ressources et l'alignement stratégique.
- Amélioration continue : pour chaque résultat, établir des mécanismes d'évaluation et de perfectionnement continus. Cela garantit que les efforts de modernisation restent adaptatifs et réactifs à l'évolution des paysages technologiques et des besoins commerciaux.

Voici une discussion détaillée de chaque résultat visé :

Architecture cible

- Définition : L'évaluation permet de définir une architecture cible claire et évolutive pour les charges de travail génératives liées à l'IA.
- Composants : Cela inclut la sélection des services cloud appropriés, la conception de pipelines de données et la garantie de l'interopérabilité du système.
- Avantages : Une architecture bien définie favorise l'évolutivité, la fiabilité et l'optimisation des performances, et fournit une base solide pour la modernisation.

Préparation du client

- Évaluation : Évaluez l'état actuel de l'infrastructure, des processus et de la culture de l'organisation afin de déterminer si elle est prête à adopter la modernisation de l'IA générative.
- Critères : Cela implique d'évaluer les capacités techniques, la qualité des données et la volonté organisationnelle d'accepter le changement.
- Résultat : L'identification des lacunes et des domaines à améliorer garantit que l'organisation est prête à effectuer une transition harmonieuse vers des solutions et des technologies modernes.

Utilisez les objectifs du cas et étirez les objectifs

- Les objectifs du cas d'utilisation établissent des objectifs clairs pour la mise en œuvre de la solution cible, en se concentrant sur des problèmes ou des opportunités commerciaux spécifiques.

  Un objectif de cas d'utilisation dans le contexte de la modernisation de l'IA générative fait référence à un objectif spécifique et mesurable qu'une organisation vise à atteindre en mettant en œuvre des solutions d'IA générative. Ces objectifs sont généralement alignés sur des objectifs commerciaux plus généraux et visent à relever des défis ou à saisir des opportunités spécifiques au sein de l'organisation. Voici des exemples d'objectifs de cas d'utilisation :

  - Réduire le temps de réponse du service client de 50 % en utilisant des chatbots génératifs alimentés par l'IA.
  - Améliorer l'efficacité de la révision du code de 30 % grâce à une analyse de code générative assistée par l'IA.
  - Améliorer la précision de la détection des fraudes de 25 % grâce à la reconnaissance générative des formes par IA.
- Les objectifs ambitieux définissent des cibles ambitieuses qui repoussent les limites de ce que la modernisation de l'IA générative peut atteindre au sein de l'organisation.
- Impact : La définition d'objectifs à la fois réalisables et ambitieux permet d'aligner les initiatives de modernisation de l'IA générative sur les objectifs commerciaux stratégiques et d'encourager l'innovation.

Estimation de l'effort

- Objectif : Une estimation précise des efforts facilite la planification des ressources et garantit que les projets sont livrés dans les délais et dans les limites du budget.
- Portée : Estimez les ressources, le temps et le budget nécessaires pour mettre en œuvre le plan de modernisation de l'IA générative.

- Facteurs : Tenez compte de la complexité technique, des défis d'intégration et des risques potentiels.

Besoins en matière d'habilitation

- Formation et développement : identifiez les compétences et les connaissances requises pour une adoption réussie de la modernisation de l'IA générative.
- Ressources : Déterminer le besoin de programmes de formation, d'ateliers et d'autres activités d'habilitation.
- Résultat : Veiller à ce que le personnel possède les compétences nécessaires améliore l'efficacité des initiatives de modernisation de l'IA générative et favorise le succès à long terme.

Plan de mise en œuvre

- Feuille de route : Élaborez un plan détaillé décrivant les étapes nécessaires à la modernisation de l'IA générative.
- Jalons : définissez les principaux jalons et livrables pour suivre les progrès.
- Avantages : Un plan de mise en œuvre clair fournit une orientation et une responsabilisation, et facilite une approche structurée de la modernisation de l'IA générative.

Considérations relatives à l'évaluation et conditions préalables

Commencez par des cas d'utilisation clairs

Identifiez les problèmes ou opportunités commerciaux spécifiques auxquels l'IA générative peut répondre. Concentrez-vous sur les cas d'utilisation qui correspondent aux objectifs commerciaux stratégiques et offrent des avantages mesurables. Priorisez les cas d'utilisation qui ciblent les défis courants au sein de l'organisation afin de garantir que l'architecture de la solution puisse servir de modèle pour plusieurs scénarios.

Lancer le processus d'évaluation avec une compréhension générale des applications potentielles de l'IA générative est bénéfique mais pas obligatoire. Le questionnaire inclus dans ce guide répond à différents niveaux de préparation, qu'il s'agisse d'organisations dont les cas d'utilisation sont bien définis ou de celles qui n'ont que des idées générales. Le processus d'évaluation sert à :

- Affinez et clarifiez ces idées de cas d'utilisation initiaux.
- Identifiez de nouveaux cas d'utilisation potentiels.
- Définissez des objectifs spécifiques et mesurables pour chaque cas d'utilisation.
- Évaluez la faisabilité et l'impact potentiel de chaque cas d'utilisation.

Prenons un exemple hypothétique : une société de services financiers décide d'explorer la modernisation de l'IA générative. Ils commencent par une idée générale de l'amélioration de leur service client et de leurs processus de détection des fraudes.

- Évaluation initiale : Le questionnaire les aide à évaluer leurs systèmes actuels, la qualité des données et l'état de préparation de l'organisation à l'adoption de l'IA générative.
- Affinement des cas d'utilisation : grâce au processus d'évaluation, ils affinent leurs idées initiales en deux cas d'utilisation spécifiques :
  - Implémentation d'un chatbot génératif alimenté par l'IA pour les demandes des clients
  - Utilisation de l'IA générative pour détecter les fraudes transactionnelles en temps réel
- Définition des objectifs : pour chaque cas d'utilisation, ils définissent des objectifs spécifiques :
  - Réduire le temps de réponse du service client de 40 % en 6 mois

    - Améliorez la précision de la détection des fraudes de 20 % et réduisez les faux positifs de 15 %
- Des objectifs ambitieux : ils ont également fixé ces objectifs ambitieux :
    - Obtenez 80 % de satisfaction client grâce aux réponses assistées par l'IA
    - Développez un modèle de détection prédictive des fraudes qui identifie les nouveaux modèles de fraude
- Définition du MVP : le questionnaire les aide à déterminer un MVP pour chaque cas d'utilisation, en se concentrant sur les fonctionnalités essentielles qui apportent une valeur immédiate.
- Architecture cible : Enfin, ils développent une architecture cible qui prend en charge un ou les deux cas d'utilisation et garantit l'évolutivité et l'intégration avec les systèmes existants.

Garantir l'alignement des activités

Alignez les initiatives d'IA générative sur la stratégie et les objectifs commerciaux globaux. Pour chaque cas d'utilisation, élaborez une proposition de valeur claire qui montre comment l'IA générative contribue à la croissance, à l'efficacité ou à l'innovation des entreprises. Établissez des mesures pour mesurer l'impact des mises en œuvre de l'IA générative sur les indicateurs de performance clés (KPIs).

Mettre en œuvre la gouvernance et la supervision

Créez un comité directeur interfonctionnel chargé de superviser les initiatives en matière d'IA générative. Élaborer des politiques et des directives pour une utilisation responsable de l'IA, en tenant compte des considérations éthiques et des biais potentiels. Établissez un processus d'examen des projets d'IA générative afin de garantir la conformité aux normes organisationnelles et aux exigences réglementaires.

Données d'adresse et prérequis techniques

Évaluez et améliorez la qualité des données, et mettez en œuvre des pratiques de gouvernance des données afin de garantir des entrées fiables pour les modèles d'IA génératifs. Développez une stratégie de données qui aborde la collecte, le stockage et la gestion des données spécifiques aux besoins en IA générative. Évaluez et améliorez l'infrastructure de données pour prendre en charge le volume et la rapidité des données nécessaires aux charges de travail génératives liées à l'IA.

Tenez compte des besoins en ressources informatiques

Évaluez l'infrastructure informatique actuelle et identifiez les lacunes en matière de capacité de calcul pour les charges de travail génératives liées à l'IA. Planifiez des ressources informatiques évolutives, en envisageant des options telles que les services cloud ou les clusters de calcul hautes performances sur site. Optimisez l'allocation des ressources pour équilibrer les performances et la rentabilité des charges de travail de formation et d'inférence.

Aborder les implications en matière de confidentialité et de sécurité

Mettez en œuvre des mesures de sécurité robustes pour protéger les données sensibles utilisées dans le cadre de la formation et des opérations liées à l'IA générative. Garantissez le respect des réglementations en matière de protection des données telles que le règlement général sur la protection des données (RGPD) ou la loi californienne sur la protection de la vie privée des consommateurs (CCPA) lors du traitement des informations personnelles. Développez des protocoles pour le déploiement et la surveillance sécurisés des modèles afin d'empêcher tout accès non autorisé ou toute utilisation abusive des capacités d'IA générative.

Impliquer les parties prenantes dès

Impliquez les principales parties prenantes dès le début pour obtenir l'adhésion et le soutien des dirigeants. Communiquez clairement les avantages et l'impact potentiel des initiatives de modernisation, en particulier pour les charges de travail génératives liées à l'IA. Fournir des formations et des ressources pour aider les parties prenantes à comprendre les technologies d'IA générative et leurs implications.

Itérer et apprendre

Adoptez une approche progressive qui vous permet d'affiner les solutions cibles. Utilisez des boucles de rétroaction pour améliorer en permanence l'architecture et les processus de charge de travail. Évaluez régulièrement les performances et l'impact des mises en œuvre de l'IA générative, et ajustez les stratégies selon les besoins en fonction des résultats concrets et de l'évolution des besoins de l'entreprise.

Questionnaire d'évaluation de la charge de travail de

Les sections suivantes proposent des questions que vous pouvez utiliser pour évaluer les différents aspects de la modernisation des charges de travail génératives liées à l'IA pour votre organisation. Ce questionnaire complet évalue le degré de préparation de votre entreprise à adopter et à mettre en œuvre des charges de travail basées sur l'IA générative en posant des questions portant sur des domaines clés, notamment les cas d'utilisation, l'architecture, le stockage, la conformité, l'intégration, les tests, le déploiement et la stratégie en matière de données. En abordant les aspects critiques de la mise en œuvre de l'IA générative, de l'infrastructure technique aux considérations réglementaires, ce questionnaire vous aide à identifier les forces, les lacunes et les opportunités de votre parcours de modernisation de l'IA.

Rubriques :

- État de préparation
- Cas d’utilisation
- Architecture
- Stockage
- Réglementation et conformité
- Integration
- Test
- Déploiement et automatisation
- Stratégie en matière de données

Vous pouvez également télécharger le questionnaire au format Microsoft Excel et l'utiliser pour enregistrer vos informations.

Téléchle questionnaire

État de préparation

Question	Exemple de réponse
Disposez-vous de AWS comptes qui peuvent être utilisés pour ces charges de travail ?	Oui ou non
Avez-vous déjà conclu un accord d'entreprise avec AWS ?	Oui ou non
Dans quelle mesure votre infrastructure cloud actuelle est-elle évolutive pour gérer les charges de travail génératives liées à l'IA ?	Notre infrastructure cloud est hautement évolutive, avec des capacités de mise à l'échelle automatique pour les ressources informatiques et les systèmes de stockage distribués conçus pour gérer efficacement les charges de travail d'IA générative à grande échelle.
Disposez-vous de capacités de pipeline de données pour le prétraitement et l'ingénierie des fonctionnalités à grande échelle ?	Nos pipelines de données utilisent des frameworks de traitement distribués tels qu'Apache Spark pour le prétraitement des données à grande échelle et l'ingénierie des fonctionnalités, avec un support pour le traitement des données par lots et en streaming .
Disposez-vous de capacités de provision nement et de gestion de comptes ?	Oui ou non
Comment décririez-vous les connaissances de votre organisation en matière d'IA et sa volonté d'adopter les technologies d'IA générative ?	Notre organisation a beaucoup investi dans des programmes de formation en IA, et la plupart du personnel technique a suivi une formation de base sur l'IA et le ML. L'organis ation possède une culture de l'innovation qui intègre les nouvelles technologies, notamment l'IA générative.

Question	Exemple de réponse
Quelle est l'expertise en matière d'intelligence artificielle et de machine learning existant au sein de votre organisation, et comment est-elle distribuée ?	Nous disposons d'un centre d'excellence dédié à l'IA composé de data scientists et d'ingénie urs ML expérimentés. Nous améliorons les compétences des experts du domaine de différentes unités commerciales afin qu'ils maîtrisent l'IA et identifient les cas d'utilisation de l'IA générative.
Avez-vous une analyse de rentabilisation de haut niveau qui expose les objectifs, les avantages et les coûts du programme cloud ?	Oui ou non
Quel est votre calendrier pour mettre la solution en production ?	Des semaines, des mois, etc.
Un engagement financier a-t-il été pris par vos principales parties prenantes (par exemple, CFO, CIT/CTO, COO) ?	Oui ou non
Comment garantissez-vous le respect des réglementations en matière de protection des données dans le cadre de vos initiatives d'IA générative ?	Nous disposons d'une équipe dédiée à la conformité qui travaille en étroite collabora tion avec nos équipes d'IA. Nous réalisons régulièrement des évaluations de l'impact sur la vie privée, mettons en œuvre des principes de protection des données dès la conception et conservons des dossiers de traitement des données détaillés pour tous les projets d'IA générative.
Quel est le degré de maturité de vos systèmes existants qui s'intègrent aux nouvelles technologies d'IA générative ?	Notre architecture informatique est basée sur des microservices et APIs permet une intégrati on flexible des nouvelles technologies génératives d'IA. Ces systèmes sont normalisés selon des formats de données et des protocoles communs afin de garantir l'interopérabilité.

Question	Exemple de réponse
Quelle expérience avez-vous en matière d'opérationnalisation de modèles de machine learning, et comment cela pourrait-il s'appliquer aux systèmes d'IA générative ?	Nous avons établi MLOps des pratiques, notamment des pipelines de déploiement de modèles automatisés, des systèmes de surveillance et des cadres de tests A/B. Ces pratiques sont en cours d'adaptation pour répondre aux exigences uniques des modèles d'IA générative à grande échelle.

Cas d’utilisation

Question	Exemple de réponse
Quel est l'objectif principal ou le critère de réussite du cas d'utilisation ?	Pour améliorer le temps de réponse du support client, augmenter le taux de conversion des ventes, améliorer les recommandations de produits. Également : pour améliorer la satisfact ion des utilisateurs, le taux d'achèvement des tâches, la qualité des réponses, etc.
Comment ce cas d'utilisation s'aligne-t-il sur les objectifs stratégiques de votre organisation ?	Cela correspond à notre objectif stratégiq ue d'améliorer la satisfaction des clients en réduisant les temps de réponse du service client.
Quel est le volume de données ou de demandes attendu pour le cas d'utilisation ?	500 transactions par seconde (TPS).
Quels types de sources de données sont nécessaires pour prendre en charge vos charges de travail génératives liées à l'IA ?	Bases de données structurées internes (dossiers clients, données de vente, etc.) ; données textuelles non structurées provenant de documents, d'e-mails et de réseaux sociaux ; fichiers audio et vidéo pour les tâches de reconnaissance vocale et d'image ; données de streaming en temps réel provenant d'apparei

Question	Exemple de réponse
	ls et de capteurs IoT ; ensembles de données publics et à des APIs fins d'enrichissement.
À quelle fréquence devez-vous mettre à jour ou actualiser les données provenant de ces sources ?	Bases de données transactionnelles : mises à jour en temps quasi réel ; référentiels de documents : mises à jour quotidiennes par lots ; flux de réseaux sociaux : mises à jour horaires ; données des capteurs IoT : diffusion continue en temps réel ; ensembles de données publics : mises à jour mensuelles ou trimestrielles.
Quels formats de données vos modèles d'IA générative ont-ils besoin en entrée ?	Données structurées : tables de base de données CSV, JSON et SQL ; données texte : texte brut, PDF et HTML ; données d'image : JPEG, PNG et TIFF ; données audio : WAV et MP3 ; données vidéo : MP4 et AVI.
Quelles sont vos principales préoccupations en matière de qualité des données pour les charges de travail génératives liées à l'IA ?	Exhaustivité : garantir qu'aucun champ critique n'est manquant ; exactitude : vérification de l'exactitude des données et élimination des erreurs ; cohérence : maintien de formats et de valeurs uniformes entre les sources ; actualité : garantie que les données sont à jour pour une inférence en temps réel ; pertinence : confirmat ion que les données correspondent à la tâche spécifique d'IA générative.
Quelles sont les principales exigences de performance (par exemple, temps de réponse, débit, précision) ?	Précision de 95 % ; temps de réponse inférieur à 500 ms ; capacité à traiter 1 000 requêtes/ sec. Haute précision (95% +), précision modérée (80-90%), meilleur effort, etc.
En avez-vous un autre KPIs pour mesurer le succès de ce cas d'utilisation ?	KPIs Les éléments clés incluent la réduction du taux d'erreur, le gain de temps par transaction et les scores de satisfaction client.

Question	Exemple de réponse
Quel est le niveau de précision du modèle souhaité, et quel est son équilibre avec le coût ?	Haute précision (> 90%) à coût modéré, précision modérée (70-80%) à faible coût, etc.
Quels sont les principaux cas d'utilisation ou scénarios de la solution d'IA générative ?	Chatbot du service client, génération de contenu, recommandation de produits, etc.
Quels sont les utilisateurs ou les personnalités cibles du système d'IA générative ?	Agents du service client, équipe marketing, employés, utilisateurs finaux, etc.
Quel est le volume de demandes ou d'utilisa teurs attendu ?	1 000 demandes par jour ; 10 000 utilisateurs actifs par mois.
Existe-t-il des contraintes ou des exigences spécifiques liées aux cas d'utilisation ?	Réponse en temps réel, support multilingue, confidentialité des données, etc.
Disposez-vous d'un budget alloué au développement et à la maintenance de la solution d'IA générative ?	Le coût de développement initial est estimé à 200 000$, avec des coûts de maintenance annuels de 50 000$.
Quels sont le retour sur investissement (ROI) et la période de remboursement prévus pour ce cas d'utilisation ?	Retour sur investissement attendu de 150 % sur trois ans, avec une période de rembourse ment de 18 mois.
Y a-t-il des coûts cachés ou des économies potentielles à envisager ?	Les économies potentielles incluent la réduction des coûts des heures supplémen taires. Les coûts cachés peuvent impliquer une formation supplémentaire pour le personnel.
Quelles sont l'évolutivité et les futures possibili tés d'extension de cette solution d'IA générativ e ?	La solution est conçue pour s'adapter à nos opérations, avec la possibilité de l'étendre à d'autres départements à l'avenir.
Comment garantir l'équité et atténuer les biais dans vos modèles d'IA générative ?	Nous prévoyons d'atténuer les biais grâce à une collecte de données diversifiée, à des audits réguliers des biais et à la mise en œuvre de techniques d'atténuation des biais.

Question	Exemple de réponse
Quels processus avez-vous mis en place pour répondre aux préoccupations éthiques ou aux conséquences imprévues ?	Nous gérerons les préoccupations éthiques grâce à un plan établi de réponse aux incidents liés à l'IA, à des évaluations régulières des risques éthiques, à un système de signaleme nt anonyme pour les employés, à la collabora tion avec des experts externes en éthique, ainsi qu'à un suivi et à un ajustement continus des modèles déployés en fonction des commentai res.
Comment abordez-vous la hiérarchisation et le séquençage des évaluations de la charge de travail générative de l'IA dans les différents projets et départements de votre organisation ?	En menant une enquête de haut niveau auprès de tous les départements afin d'identifier les cas d'utilisation potentiels de l'IA générative et en les évaluant sur la base de trois critères clés : impact commercial, faisabilité technique et considérations éthiques. Les projets ayant un impact potentiel élevé, des obstacles technique s réduits et des préoccupations éthiques minimales sont prioritaires.

Architecture

Question	Exemple de réponse
Quel type de modèle ou d'architecture d'IA générative est envisagé ?	Transformateur, réseau neuronal convoluti f (CNN), réseau neuronal récurrent (RNN), arbres de décision, etc.
Quelle est l'échelle ou le volume de données et de calculs attendus ?	Des millions d'utilisateurs, des pétaoctets de données, etc.
Quelles sont les exigences matérielles (par exemple, CPUs ou GPUs) pour la formation et l'inférence ?	Clusters de processeurs haut GPUs de gamme, instances cloud, etc.

Question	Exemple de réponse
Comment le modèle d'IA générative sera-t-il mis à jour ou réentraîné au fil du temps ?	Grâce à un apprentissage continu, à une reconversion périodique, à des mises à jour manuelles, etc.
Quelles sont les exigences en matière de prétraitement des données et d'ingénierie des fonctionnalités ?	Nettoyage du texte, augmentation de l'image, sélection de fonctionnalités, etc.
Comment le système d'IA générative gérera-t-il les cas extrêmes, les valeurs aberrantes ou les entrées peu fiables ?	En recourant à une supervision humaine, en demandant des éclaircissements, etc.
Quelles sont les exigences de latence pour l'application d'IA générative ?	Traitement par lots en temps réel, en temps quasi réel, etc.

Stockage

Question	Exemple de réponse
Où seront stockées les données d'entraîn ement ?	Dans le stockage dans le cloud (par exemple, Amazon S3, le stockage de fichiers, le stockage par blocs ou le stockage d'objets), dans le stockage sur site, etc.
Quelles sont les exigences de stockage pour les données de formation et les artefacts du modèle (par exemple, capacité, durabilité, disponibilité) ?	Stockage à l'échelle du pétaoctet, durabilité élevée (durabilité de 99,999999999 %), haute disponibilité, etc.
Quelles sont les exigences en matière de conservation et de sauvegarde des données de formation et des artefacts du modèle ?	Conservation des données pendant x ans, sauvegardes quotidiennes, sauvegardes hors site, etc.
Quels formats de fichiers sont principal ement utilisés pour stocker vos ensembles	Fichiers Parquet pour les données structurées, HDF5 les grands tableaux multidimensionnels

Question	Exemple de réponse
de données de formation basés sur l'IA (par exemple, CSV, JSON, Parquet HDF5) ?	et les données non structurées telles que les images et le texte. Nous utilisons des formats spécialisés, par exemple TFRecord pour optimiser le chargement des données pendant l'entraînement.
Comment sont organisés vos ensembles de données d'entraînement : sous forme de fichiers individuels, dans des bases de données ou à l'aide de formats de données d'IA spécialis és ?	Les ensembles de données de petite à moyenne taille sont stockés sous forme de fichiers Parquet individuels dans le stockage d'objets pour plus de flexibilité. Les grands ensembles de données sont stockés dans une base de données distribuée (Cassandra) pour gérer l'échelle.
Utilisez-vous des techniques de compression ou d'encodage de données spécifiques aux données d'entraînement génératives issues de l'IA ?	Pour les données tabulaires, nous utilisons des techniques de codage par dictionnaire et de compression de bits disponibles dans Parquet. Pour les images, nous utilisons une compressi on JPEG avec perte avec des paramètres de qualité optimisés pour nos modèles.
Comment gérez-vous le versionnement et le stockage des différentes itérations d'ensembles de données de formation ? Quel impact cela a-t-il sur l'ensemble de vos besoins de stockage ?	Nous utilisons un système de version des données (DVC) intégré à notre plateforme ML.

Réglementation et conformité

Question	Exemple de réponse
Quelles sont les réglementations ou exigences de conformité pertinentes pour la solution d'IA générative (par exemple, RGPD, HIPAA, PCI-DSS) ?	RGPD pour le traitement des données personnelles, HIPAA pour les données de santé, PCI-DSS pour les données de paiement, etc.

Question	Exemple de réponse
Quels directives ou cadres éthiques en matière d'IA générative votre organisation a-t-elle adoptés ?	Nous avons mis en œuvre nos propres directives en matière d'IA responsable. Tous les projets d'IA générative font l'objet d'un examen éthique avant d'être approuvés et déployés.
Quelles sont les exigences de sécurité pour le système d'IA générative ?	Chiffrement des données, communication réseau sécurisée, audits de sécurité réguliers.
Quelles sont les exigences en matière de confidentialité et de protection des données ?	Anonymisation des données, chiffrement, contrôle d'accès, etc.
Quelles sont les exigences requises pour que la solution gère des données sensibles ou confidentielles ?	Contrôles d'accès stricts, masquage des données, exigences de résidence des données, etc.
Comment seront gérées l'authentification et l'autorisation des utilisateurs ?	En utilisant des clés d'API OAuth, une authentif ication unique (SSO) et un contrôle d'accès basé sur les rôles (RBAC).
Comment la solution sera-t-elle surveillée et gérée en production ?	En utilisant des outils de surveillance tels que Prometheus et Datadog, des outils de journalis ation tels que ELK Stack, des systèmes d'alerte, etc.

Integration

Question	Exemple de réponse
Quelles sont les exigences pour intégrer la solution d'IA générative aux systèmes ou sources de données existants ?	REST APIs, files d'attente de messages, connecteurs de base de données, etc.

Question	Exemple de réponse
Comment les données seront-elles ingérées et prétraitées pour la solution d'IA générative ?	En utilisant le traitement par lots, le streaming de données, les transformations de données et l'ingénierie des fonctionnalités.
Comment les résultats de la solution d'IA générative seront-ils consommés ou intégrés aux systèmes en aval ?	Par le biais de points de terminaison d'API, de files d'attente de messages, de mises à jour de bases de données, etc.
Quels modèles d'intégration pilotés par les événements peuvent être utilisés pour la solution d'IA générative ?	Les files d'attente de messages (comme Amazon SQS, Apache Kafka, RabbitMQ), les systèmes pub/sub, les webhooks, les plateforms de streaming d'événements.
Quelles approches d'intégration basées sur les API peuvent être utilisées pour connecter la solution d'IA générative à d'autres systèmes ?	RESTful APIs, GraphQL APIs, SOAP APIs (pour les anciens systèmes).
Quels composants de l'architecture de microservices peuvent être utilisés pour l'intégration d'une solution d'IA générative ?	Maillage de services pour la communication interservices, les passerelles d'API, l'orchestration de conteneurs (par exemple, Kubernetes).
Comment l'intégration hybride peut-elle être mise en œuvre pour la solution d'IA générative ?	En combinant des modèles pilotés par les événements pour les mises à jour en temps réel, le traitement par lots des données historiques et APIs pour l'intégration de systèmes externes.
Comment les résultats de la solution d'IA générative peuvent-ils être intégrés aux systèmes en aval ?	Par le biais de points de terminaison d'API, de files d'attente de messages, de mises à jour de bases de données, de webhooks et d'exportations de fichiers.

Question	Exemple de réponse
Quelles mesures de sécurité doivent être prises en compte pour intégrer la solution d'IA générative ?	Mécanismes d'authentification (tels que OAuth JWT), chiffrement (en transit et au repos), limitation du débit des API et listes de contrôle d'accès (ACLs).
Comment prévoyez-vous d'intégrer des frameworks open source tels que LlamaIndex ou LangChain dans votre pipeline de données existant et votre flux de travail d'IA générative ?	Nous prévoyons de l'utiliser LangChain pour créer des applications d'IA génératives complexes, en particulier pour ses capacités de gestion des agents et de la mémoire. Notre objectif est de faire en sorte que 60 % de nos projets d'IA générative LangChain soient utilisés au cours des 6 prochains mois.
Comment garantirez-vous la compatibilité entre les frameworks open source que vous avez choisis et votre infrastructure de données existante ?	Nous sommes en train de créer une équipe d'intégration dédiée pour garantir une compatibilité fluide. D'ici le troisième trimestre , notre objectif est de disposer d'un pipeline entièrement intégré permettant d'indexer et de récupérer efficacement les données au sein de notre structure de lac de données actuelle. LlamaIndex
Comment comptez-vous tirer parti des composants modulaires des frameworks, par exemple LangChain pour le prototypage rapide et l'expérimentation ?	Nous mettons en place un environnement sandbox dans lequel les développeurs peuvent rapidement prototyper en utilisant LangChain les composants.
Quelle est votre stratégie pour suivre les mises à jour et les nouvelles fonctionnalités de ces frameworks open source en évolution rapide ?	Nous avons chargé une équipe de surveiller les GitHub référentiels et les forums communaut aires pour LangChain et LlamaIndex. Nous prévoyons d'évaluer et d'intégrer des mises à jour majeures tous les trimestres, en mettant l'accent sur l'amélioration des performances et les nouvelles fonctionnalités.

Test

Question	Exemple de réponse
Quelles sont les exigences en matière de tests (par exemple, tests unitaires, tests d'intégration, end-to-end tests) ?	Tests unitaires pour des composants individuels, tests d'intégration avec des systèmes externes, end-to-end tests pour des scénarios critiques, etc.
Comment garantissez-vous la qualité et la cohérence des données entre les différentes sources pour la formation à l'IA générative ?	Nous maintenons la qualité des données grâce à des outils automatisés de profilage des données, à des audits de données réguliers et à un catalogue de données centralisé. Nous avons mis en place des politiques de gouvernance des données afin de garantir la cohérence entre les sources et de maintenir le lignage des données.
Comment le modèle d'IA générative sera-t-il évalué et validé ?	En utilisant un ensemble de données résistant, une évaluation humaine, des tests A/B, etc.
Quels sont les critères d'évaluation des performances et de la précision du modèle d'IA générative ?	Précision, rappel, score F1, perplexité, évaluation humaine, etc.
Comment les cas extrêmes et les cas critiques seront-ils identifiés et traités ?	En utilisant une suite de tests complète, une évaluation humaine, des tests contradictoires, etc.
Comment testerez-vous les biais potentiels dans le modèle d'IA générative ?	En utilisant une analyse de parité démographique, des tests d'égalité des chances, des techniques de débiais contradictoires, des tests contrefactuels, etc.
Quels paramètres seront utilisés pour mesurer l'équité des résultats du modèle ?	Rapport d'impact disparate, chances égalisées, parité démographique, indicateurs d'équité individuels, etc.

Question	Exemple de réponse
Comment garantirez-vous une représentation diversifiée dans vos ensembles de données de test pour la détection des biais ?	En utilisant un échantillonnage stratifié par groupes démographiques, en collaboration avec des experts en diversité, en utilisant des données synthétiques pour combler les lacunes, etc.
Quel processus sera mis en œuvre pour le suivi continu de l'équité des modèles après le déploiement ?	Audits d'équité réguliers, systèmes automatisés de détection des biais, analyse des commentai res des utilisateurs, formation continue périodique avec des ensembles de données mis à jour, etc.
Comment aborderez-vous les biais intersect ionnels dans le modèle d'IA générative ?	En utilisant une analyse d'équité intersect ionnelle, des tests en sous-groupes, une collaboration avec des experts du domaine sur l'intersectionnalité, etc.
Comment testerez-vous les performances du modèle dans différents contextes linguistiques et culturels ?	En utilisant des ensembles de tests multiling ues, une collaboration avec des experts culturels, des mesures d'équité localisées, des études comparatives interculturelles, etc.

Déploiement et automatisation

Question	Exemple de réponse
Quelles sont les exigences en matière de mise à l'échelle et d'équilibrage de charge ?	Routage intelligent des demandes ; système de mise à l'échelle automatique ; optimisation pour les démarrages à froid rapides en utilisant des techniques telles que la mise en cache des modèles, le chargement différé et les systèmes de stockage distribués ; conception du système pour gérer des modèles de trafic rapides et imprévisibles.

Question	Exemple de réponse
Quelles sont les exigences relatives à la mise à jour et au déploiement de nouvelles versions ?	Déploiements bleu/vert, versions de Canary, mises à jour continues, etc.
Quelles sont les exigences en matière de reprise après sinistre et de continuité des activités ?	Procédures de sauvegarde et de restauration, mécanismes de basculement, configurations de haute disponibilité, etc.
Quelles sont les exigences pour automatiser la formation, le déploiement et la gestion du modèle d'IA générative ?	Pipeline de formation automatisé, déploiement continu, mise à l'échelle automatique, etc.
Comment le modèle d'IA générative sera-t-il mis à jour et réentraîné au fur et à mesure que de nouvelles données seront disponibles ?	Par le biais d'une reconversion périodique, d'un apprentissage progressif, d'un apprentissage par transfert, etc.
Quelles sont les exigences relatives à l'automatisation de la surveillance et de la gestion ?	Alertes automatisées, mise à l'échelle automatique, autoréparation, etc.
Quel est votre environnement de déploiement préféré pour les charges de travail génératives liées à l'IA ?	Une approche hybride qui utilise AWS pour la formation des modèles et notre infrastructure sur site pour l'inférence afin de répondre aux exigences de résidence des données.
Y a-t-il des plateformes cloud spécifiques que vous préférez pour les déploiements d'IA générative ?	Services AWS, en particulier Amazon SageMaker AI pour le développement et le déploiement de modèles, et Amazon Bedrock pour les modèles de base.
Quelles technologies de conteneurisation envisagez-vous pour les charges de travail génératives liées à l'IA ?	Nous voulons standardiser les conteneurs Docker orchestrés avec Kubernetes afin de garantir la portabilité et l'évolutivité dans notre environnement hybride.
Avez-vous des outils préférés pour le CI/CD dans votre pipeline d'IA générative ?	GitLab pour le contrôle de version et les pipelines CI/CD, intégrés à Jenkins pour les tests et les déploiements automatisés.

Question	Exemple de réponse
Quels outils d'orchestration envisagez-vous pour gérer les flux de travail d'IA générative ?	Apache Airflow pour l'orchestration des flux de travail, en particulier pour le prétraitement des données et les pipelines d'entraînement des modèles.
Avez-vous des exigences spécifiques en matière d'infrastructure sur site pour prendre en charge les charges de travail génératives liées à l'IA ?	Nous investissons dans des serveurs accélérés par GPU et dans des réseaux haut débit pour prendre en charge les charges de travail d'inférence sur site.
Comment prévoyez-vous de gérer le versionnement et le déploiement des modèles dans différents environnements ?	Nous prévoyons de l'utiliser MLflow pour le suivi des modèles et le versionnement, et de l'intégrer à notre infrastructure Kubernetes pour un déploiement fluide dans tous les environnements.
Quels outils de surveillance et d'observabilité envisagez-vous pour les déploiements d'IA générative ?	Prometheus pour la collecte des métriques et Grafana pour la visualisation, avec des solutions de journalisation personnalisées supplémentaires pour la surveillance spécifique au modèle.
Comment abordez-vous le mouvement et la synchronisation des données dans un modèle de déploiement hybride ?	Nous utiliserons AWS DataSync un transfert de données efficace entre le stockage sur site et AWS des tâches de synchronisation automatisées planifiées en fonction de nos cycles de formation.
Quelles mesures de sécurité mettez-vous en œuvre pour les déploiements d'IA générative dans différents environnements ?	Nous utiliserons l'IAM pour les ressources cloud, intégrées à notre Active Directory sur site pour implémenter le end-to-end chiffrement et la segmentation du réseau afin de sécuriser les flux de données.

Stratégie en matière de données

Question	Exemple de réponse
Quels types de données spécifiques sont essentiels pour vos charges de travail génératives liées à l'IA, et quel pourcentage d'entre elles sont actuellement accessibles ?	Les journaux d'appels des clients et les données relatives aux avis sur les produits sont essentiels. Actuellement, 85 % de ces types de données sont accessibles pour nos projets d'IA générative.
Comment garantissez-vous et mesurez-vous la qualité de vos données ?	Nous avons mis en place des mesures de qualité des données, notamment l'exhaustivité, l'exactitude, la cohérence et l'actualité. Nous utilisons des outils automatisés pour évaluer régulièrement ces indicateurs et disposons d'une équipe dédiée au nettoyage et à l'enrichissement des données.
Quel pourcentage de vos données répond à vos normes de qualité en matière d'utilisation de l'IA générative ?	Actuellement, 78 % de nos données répondent à nos normes de qualité. Nous visons 95 % au cours des 12 prochains mois grâce à de meilleurs processus de nettoyage des données.
Comment comptez-vous renforcer la confiance de vos parties prenantes quant à l'utilisation des données dans l'IA générative ?	Nous mettons en place un comité d'éthique de l'IA, fournissant des explications claires sur les décisions relatives à l'IA et menant des audits trimestriels sur l'IA pour garantir la transparence et l'équité.
Dans quelle mesure votre documentation relative aux sources de données et au lignage est-elle complète ?	Nous maintenons un catalogue de données détaillé qui inclut les métadonnées de toutes nos sources de données, y compris l'origine, la fréquence des mises à jour et l'utilisation. Nous utilisons des outils de traçabilité des données pour suivre la manière dont les données circulent et se transforment dans nos systèmes.

Question	Exemple de réponse
Comment garantissez-vous la diversité de vos ensembles de données afin d'éviter les biais dans les modèles d'IA ?	Nous nous approvisionnons activement en données provenant de divers groupes démographiques et auditons régulièrement nos ensembles de données pour détecter tout biais représentatif. Nous utilisons également des techniques de génération de données synthétiq ues pour équilibrer les catégories sous-repr ésentées.
Quel est votre taux de rafraîchissement des données pour les modèles d'IA générative critiques, et comment déterminez-vous cette fréquence ?	Les modèles critiques sont actualisés chaque semaine. Cette fréquence est déterminée par les indicateurs de performance des tests A/B, et nous visons une dégradation maximale de 2 % entre les actualisations.
Combien de versions d'ensembles de données critiques maintenez-vous et pendant combien de temps ?	Nous maintenons les cinq dernières versions de chaque ensemble de données critique, avec une période de conservation de 18 mois pour chaque version.
Combien d'équipes interfonctionnelles participe nt à vos initiatives d'IA générative et ont accès à vos données ?	Nous avons trois équipes interfonctionnelles. Chaque équipe comprend des scientifiques des données, des experts du domaine, des éthiciens et des analystes commerciaux.
Quelles politiques et pratiques de gouvernance des données avez-vous mises en place ?	Nous avons un comité interfonctionnel de gouvernance des données qui supervise nos politiques en matière de données. Nous avons mis en place des contrôles d'accès basés sur les rôles, des systèmes de classification des données et des audits réguliers pour garantir le respect de notre cadre de gouvernance.

Question	Exemple de réponse
Quelles mesures avez-vous mises en place pour garantir la confidentialité des données, obtenir le consentement approprié et préserver la confidentialité ?	Nous avons mis en place un cadre complet de confidentialité des données aligné sur le RGPD et le CCPA. Cela inclut l'obtention d'un consentement explicite pour l'utilisation des données, la mise en œuvre de techniques d'anonymisation des données et des évaluatio ns régulières de l'impact sur la vie privée.
Quel pourcentage de vos ensembles de données de formation à l'IA a été audité pour détecter tout biais au cours du dernier trimestre ?	70 % de nos ensembles de données de formation sur l'IA ont été audités pour détecter tout biais au dernier trimestre. Nous mettons en œuvre des outils automatisés de détection des biais pour atteindre 100 % d'audits trimestriels.
Quelle est votre capacité actuelle de traitement des données, et dans quelle mesure prévoyez-vous avoir besoin pour les futures charges de travail d'IA générative ?	Notre capacité actuelle est de 10 TB/day. We project needing 30 TB/day en un an et nous développons notre infrastructure pour répondre à cette demande.
Quelle est votre stratégie pour trouver un équilibre entre la confidentialité des données et les besoins en données des modèles d'IA générative ?	Nous mettons en œuvre des techniques avancées d'anonymisation et de génération de données synthétiques. Notre objectif est d'augmenter nos données utilisables pour l'IA de 40 % tout en réduisant les risques de confidentialité de 60 % au cours de la prochaine année.
Quel est le pourcentage de vos ensembles de données d'apprentissage automatique (ML) étiquetés avec précision, et quel est votre taux de précision cible ?	Actuellement, 85 % de nos ensembles de données ML sont étiquetés avec précision. Nous visons un taux de précision de 95 % au cours du prochain trimestre en utilisant des techniques d'étiquetage humaines et automatis ées.

Traduire les informations issues des évaluations en résultats exploitables

Cette section fournit un cadre pour analyser les réponses au questionnaire et utiliser ces informations pour façonner l'architecture cible et les autres résultats clés de l'initiative de modernisation de l'IA générative. Ce cadre comble le fossé entre la collecte de données et la mise en œuvre, et garantit que l'évaluation oriente et oriente directement votre stratégie de modernisation.

Définition de l'architecture cible :

- Utilisez les réponses au questionnaire pour éclairer le choix des services cloud et la conception des pipelines de données.
- Assurez-vous que la conception de l'architecture prend en charge l'évolutivité et l'interopérabilité, comme indiqué dans le guide.

Évaluation de l'état de préparation du client :

- Analysez les réponses au questionnaire liées à l'infrastructure, aux processus et à la culture organisationnelle actuels.
- Identifiez les lacunes et créez un plan pour y remédier. Priorisez les lacunes essentielles au succès du MVP.

Cas d'utilisation et objectifs ambitieux :

- Extrayez les problèmes commerciaux spécifiques des réponses au questionnaire afin de définir des objectifs clairs en matière de cas d'utilisation.
- Définissez des objectifs ambitieux qui correspondent à la vision à long terme de votre organisation en matière de modernisation de l'IA générative.

Estimation de l'effort :

- Utilisez les données du questionnaire pour estimer les ressources, le temps et le budget nécessaires au MVP et à la mise en œuvre complète.
- Créez une approche progressive qui commence par le MVP et décrivez les phases suivantes.

Besoins d'habilitation :

- Sur la base des réponses au questionnaire, identifiez les lacunes en matière de compétences et les besoins de formation.
- Développez un plan de formation qui répond à la fois aux besoins immédiats des MVP et à l'adoption à long terme de l'IA générative.

Plan de mise en œuvre :

- Créez une feuille de route complète qui commence par le MVP et décrit les étapes à suivre pour une modernisation complète de l'IA générative.
- Définissez des jalons et des livrables clairs pour chaque phase de la mise en œuvre.

Étapes pratiques :

- Matrice de priorisation : créez une matrice qui fait correspondre les réponses au questionnaire aux six résultats pour aider à hiérarchiser les fonctionnalités et les efforts.
- Approche itérative : Concevez le MVP pour qu'il soit la première itération d'une série de versions planifiées, chaque version s'appuyant sur l'architecture cible complète.
- Harmonisation des parties prenantes : utilisez les résultats du questionnaire pour aligner les parties prenantes sur le champ d'application du MVP et sur l'approche progressive visant à atteindre tous les résultats.
- Boucle de feedback continue : mettez en œuvre des mécanismes pour recueillir des commentaires après le déploiement du MVP, et utilisez les informations pour affiner les plans pour les phases suivantes.
- Mise en œuvre agile : Adoptez une méthodologie agile qui permet de traiter avec flexibilité tous les résultats au fil du temps, en commençant par les résultats les plus critiques du MVP.

Étapes suivantes

Une fois que vous avez terminé l'évaluation de la charge de travail générative de l'IA, procédez comme suit :

1. Fournir une architecture cible détaillée
   - Objectif : L'architecte de solutions crée une architecture cible complète qui correspond aux objectifs de l'organisation et aux résultats de l'évaluation.
   - Composants : Cette architecture inclut la conception de l'ingestion des données, des points d'intégration et de l'interopérabilité du système afin de garantir l'évolutivité, la fiabilité et l'optimisation des performances.
2. Expliquez dans quelle mesure les spécificités du cas d'utilisation sont Services AWS adaptées
   - Cartographie des services : identifiez et cartographiez les services spécifiques Services AWS qui correspondent le mieux aux cas d'utilisation identifiés.
   - Avantages : mettez en évidence la manière dont ces services répondent aux besoins spécifiques de l'entreprise, améliorent l'efficacité et offrent une évolutivité.
3. Fournir des solutions alternatives facultatives avec des avantages et des inconvénients
   - Solutions de rechange : Présentez des solutions alternatives qui pourraient également répondre aux exigences de l'organisation.
   - Analyse : Proposez une analyse détaillée des avantages et des inconvénients de chaque alternative en tenant compte de facteurs tels que le coût, la complexité et l'alignement sur les objectifs commerciaux.
4. Fournir une estimation détaillée du prix de Services AWS
   - Analyse des coûts : fournissez une estimation détaillée des coûts pour le projet proposé Services AWS, y compris les scénarios d'utilisation potentiels et les modèles de tarification.
   - Harmonisation du budget : assurez-vous que le coût correspond aux contraintes budgétaires de l'organisation et qu'il permet de bien comprendre les implications financières.
5. Obtenez des commentaires sur l'architecture proposée
   - Engagement des parties prenantes : Interagissez avec les parties prenantes pour présenter l'architecture proposée et recueillir des commentaires.
   - Amélioration itérative : utilisez les commentaires pour affiner et améliorer la solution, et confirmer qu'elle répond aux besoins et aux attentes de toutes les parties prenantes.

FAQ

Quel est l'objectif principal de l'évaluation de la charge de travail de l'IA générative ?

L'objectif principal de l'évaluation est d'évaluer l'état de préparation d'une organisation à la modernisation de ses charges de travail génératives en matière d'IA, d'identifier les cas d'utilisation et de développer une architecture de solution cible. Il vise à définir les exigences de modernisation, à déterminer le périmètre de mise en œuvre et à préparer une modernisation réussie de l'IA générative.

Qui devrait utiliser cette évaluation ?

Cette évaluation s'adresse aux architectes de solutions, aux architectes d'entreprise et aux architectes d'applications qui souhaitent évaluer les aspects techniques de la modernisation de l'IA générative. Il est également utile aux responsables de programme et aux responsables du personnel pour évaluer l'état de préparation global, l'allocation des ressources et les besoins en matière d'habilitation.

Quels sont les principaux éléments évalués dans le cadre de l'évaluation ?

L'évaluation couvre l'état de préparation global, les cas d'utilisation, l'architecture, le stockage, les réglementations et la conformité, l'intégration, les tests, l'automatisation du déploiement et la stratégie de données. Ces composants sont essentiels pour déterminer le niveau de préparation technique et organisationnel en vue de l'adoption de la modernisation de l'IA générative.

Comment l'évaluation aide-t-elle à définir l'architecture cible ?

L'évaluation fournit une approche structurée pour évaluer les systèmes actuels et identifier les améliorations. Il vous aide à sélectionner les technologies appropriées et à concevoir des architectures évolutives conformes aux objectifs commerciaux et aux exigences des cas d'utilisation.

Quels sont les avantages d'une évaluation générative de la charge de travail liée à l'IA ?

Les avantages incluent une efficacité accrue, une meilleure prise de décision, l'assurance de la conformité, la promotion de l'innovation et la préparation à l'évolutivité. L'évaluation établit une approche stratégique de la modernisation de l'IA générative et maximise les avantages potentiels tout en atténuant les risques.

Comment les organisations peuvent-elles garantir une mise en œuvre réussie à la suite de l'évaluation ?

Organisations devraient élaborer un plan de mise en œuvre clair comprenant des étapes définies, impliquer les parties prenantes dès le début et adopter une approche itérative. La mise en place d'un centre d'excellence (CoE) et l'accent mis sur le développement des talents sont également des meilleures pratiques recommandées.

À quels défis les organisations pourraient-elles être confrontées lors de l'évaluation ?

Les défis peuvent inclure la résistance au changement, les problèmes de qualité des données et les complexités liées à la conformité. Pour relever ces défis, il faut favoriser une culture de l'innovation, garantir la disponibilité des données et mettre en œuvre des mesures de sécurité robustes.

Comment l'évaluation répond-elle aux exigences réglementaires et de conformité ?

L'évaluation évalue les mesures de conformité actuelles et identifie les lacunes. Il garantit que les solutions cibles respectent les réglementations pertinentes et les lois sur la confidentialité des données, et intègrent les meilleures pratiques de sécurité pour protéger les informations sensibles.

Quel est le rôle de l'engagement des parties prenantes dans le processus d'évaluation ?

L'engagement des parties prenantes est essentiel pour obtenir l'adhésion, aligner les initiatives de modernisation sur les objectifs commerciaux et garantir une mise en œuvre réussie. Une implication

précoce et une communication claire des avantages sont essentielles pour surmonter les résistances et favoriser le soutien.

Comment les entreprises peuvent-elles mesurer le succès de leurs initiatives de modernisation de l'IA générative après l'évaluation ?

Le succès peut être mesuré à l'aide d'indicateurs de performance clés (KPIs) qui correspondent aux objectifs commerciaux. Le suivi et l'évaluation réguliers de ces indicateurs aident à orienter la prise de décision et à démontrer la valeur de la modernisation de l'IA générative aux parties prenantes.

En quoi l'approche d'évaluation diffère-t-elle pour les organisations de différentes tailles (petites, moyennes ou entreprises) ou de différents secteurs d'activité ?

Petites organisations :

- Peut avoir des ressources et une expertise limitées pour des évaluations complètes
- Susceptible de se concentrer sur des cas d'utilisation spécifiques à fort impact plutôt que sur une adoption à l'échelle de l'entreprise
- Peut s'appuyer davantage sur des outils et services tiers pour l'évaluation
- Le processus d'évaluation pourrait être moins formel et plus souple

Entreprises de taille moyenne :

- Disposent souvent d'équipes informatiques ou de données dédiées, mais elles peuvent manquer d'expertise spécialisée en IA
- Pourrait adopter une approche progressive, en commençant par des projets pilotes dans les principaux départements
- Nécessité d'équilibrer l'innovation avec les systèmes et processus existants
- L'évaluation implique probablement des équipes interfonctionnelles

Organisations d'entreprise :

- Disposez généralement d' AI/ML équipes dédiées et de ressources supplémentaires pour une évaluation complète
- Nécessité d'envisager des intégrations complexes avec les systèmes d'entreprise existants
- Il peut y avoir des exigences réglementaires spécifiques à l'industrie à prendre en compte
- L'évaluation implique souvent des processus de gouvernance formels

Ressources

- IA générative activée AWS
- AWS propose de nouveaux guides sur l'intelligence artificielle, l'apprentissage automatique et l'IA générative pour planifier votre stratégie en matière d'IA (article de AWS blog)
- Bonnes pratiques pour créer des applications d'IA génératives AWS (article de AWS blog)
- Générateur d'applications d'IA générative activé AWS (bibliothèque de AWS solutions)
- Capacités d'IA générative (architecture AWS de référence de sécurité)
- AWS cadre des meilleures pratiques en matière d'IA générative (guide de AWS Audit Manager l'utilisateur)
- Choisir un service d'IA génératif (guide de AWS décision)
- Qu'est-ce qu'Amazon Bedrock ? (Guide de l'utilisateur d'Amazon Bedrock)
- Qu'est-ce qu'Amazon SageMaker AI ?(Guide du développeur Amazon SageMaker AI)

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	—	6 novembre 2024

AWS Glossaire des directives prescriptives

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

Nombres

7 R

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

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

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

A

ABAC

Voir contrôle d'accès basé sur les attributs.

services abstraits

Consultez la section Services gérés.

ACIDE

Voir atomicité, consistance, isolation, durabilité.

migration active-active

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

migration active-passive

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

fonction d'agrégation

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

AI

Voir intelligence artificielle.

AIOps

Voir les opérations d'intelligence artificielle.

anonymisation

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

anti-motif

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

contrôle des applications

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

portefeuille d'applications

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

intelligence artificielle (IA)

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

opérations d'intelligence artificielle (AIOps)

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

chiffrement asymétrique

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

atomicité, cohérence, isolement, durabilité (ACID)

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

contrôle d'accès par attributs (ABAC)

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

source de données faisant autorité

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

Zone de disponibilité

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

AWS Cadre d'adoption du cloud (AWS CAF)

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

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

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

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

B

mauvais bot

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

BCP

Consultez la section Planification de la continuité des activités.

graphique de comportement

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

système de poids fort

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

classification binaire

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

filtre de Bloom

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

déploiement bleu/vert

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

bot

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

botnet

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

branche

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

accès par brise-vitre

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

stratégie existante (brownfield)

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

cache de tampon

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

capacité métier

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

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

planification de la continuité des activités (BCP)

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

C

CAF

Voir le cadre d'adoption du AWS cloud.

déploiement de Canary

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

CCo E

Voir le Centre d'excellence du cloud.

CDC

Voir capture des données de modification.

capture des données de modification (CDC)

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

ingénierie du chaos

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

CI/CD

Découvrez l'intégration continue et la livraison continue.

classification

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

chiffrement côté client

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

Centre d'excellence du cloud (CCoE)

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

cloud computing

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

modèle d'exploitation du cloud

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

étapes d'adoption du cloud

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

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

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

CMDB

Consultez la base de données de gestion des configurations.

référentiel de code

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

cache passif

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

données gelées

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

vision par ordinateur (CV)

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

dérive de configuration

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

base de données de gestion des configurations (CMDB)

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

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

pack de conformité

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

intégration continue et livraison continue (CI/CD)

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

CV

Voir vision par ordinateur.

D

données au repos

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

classification des données

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

dérive des données

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

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

données en transit

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

maillage de données

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

minimisation des données

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

périmètre de données

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

prétraitement des données

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

provenance des données

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

sujet des données

Personne dont les données sont collectées et traitées.

entrepôt des données

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

langage de définition de base de données (DDL)

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

langage de manipulation de base de données (DML)

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

DDL

Voir langage de définition de base de données.

ensemble profond

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

deep learning

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

defense-in-depth

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

administrateur délégué

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

déploiement

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

environnement de développement

Voir environnement.

contrôle de détection

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

cartographie de la chaîne de valeur du développement (DVSM)

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

jumeau numérique

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

tableau des dimensions

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

catastrophe

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

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

reprise après sinistre (DR)

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

DML

Voir langage de manipulation de base de données.

conception axée sur le domaine

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

DR

Voir reprise après sinistre.

détection de dérive

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

DVSM

Voir la cartographie de la chaîne de valeur du développement.

E

EDA

Voir analyse exploratoire des données.

EDI

Voir échange de données informatisé.

informatique de périphérie

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

échange de données informatisé (EDI)

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

chiffrement

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

clé de chiffrement

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

endianisme

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

point de terminaison

Voir point de terminaison de service.

service de point de terminaison

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

planification des ressources d'entreprise (ERP)

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

chiffrement d'enveloppe

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

environment

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

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

épopée

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

ERP

Voir Planification des ressources d'entreprise.

analyse exploratoire des données (EDA)

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

F

tableau des faits

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

échouer rapidement

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

limite d'isolation des défauts

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

branche de fonctionnalités

Voir succursale.

fonctionnalités

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

importance des fonctionnalités

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

transformation de fonctionnalité

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

invitation en quelques coups

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

FGAC

Découvrez le contrôle d'accès détaillé.

contrôle d'accès détaillé (FGAC)

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

migration instantanée (flash-cut)

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

FM

Voir le modèle de fondation.

modèle de fondation (FM)

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

G

IA générative

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

blocage géographique

Voir les restrictions géographiques.

restrictions géographiques (blocage géographique)

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

Flux de travail Gitflow

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

image dorée

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

stratégie inédite

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

barrière de protection

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

H

HA

Découvrez la haute disponibilité.

migration de base de données hétérogène

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

haute disponibilité (HA)

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

modernisation des historiens

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

données de rétention

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

migration de base de données homogène

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

données chaudes

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

correctif

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

période de soins intensifs

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

I

IaC

Considérez l'infrastructure comme un code.

politique basée sur l’identité

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

application inactive

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

IIo T

Voir Internet industriel des objets.

infrastructure immuable

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

VPC entrant (d'entrée)

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

migration incrémentielle

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

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

Industry 4.0

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

infrastructure

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

infrastructure en tant que code (IaC)

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

Internet industriel des objets (IIoT)

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

VPC d'inspection

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

Internet des objets (IoT)

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

interprétabilité

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

IoT

Voir Internet des objets.

Bibliothèque d'informations informatiques (ITIL)

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

gestion des services informatiques (ITSM)

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

ITIL

Consultez la bibliothèque d'informations informatiques.

ITSM

Voir Gestion des services informatiques.

L

contrôle d'accès basé sur des étiquettes (LBAC)

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

zone de destination

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

grand modèle de langage (LLM)

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

migration de grande envergure

Migration de 300 serveurs ou plus.

LBAC

Voir contrôle d'accès basé sur des étiquettes.

principe de moindre privilège

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

lift and shift

Voir 7 Rs.

système de poids faible

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

LLM

Voir le grand modèle de langage.

environnements inférieurs

Voir environnement.

M

machine learning (ML)

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

branche principale

Voir succursale.

malware

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

services gérés

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

système d'exécution de la fabrication (MES)

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

MAP

Voir Migration Acceleration Program.

mécanisme

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

compte membre

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

MAILLES

Voir le système d'exécution de la fabrication.

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

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

microservice

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

architecture de microservices

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

Programme d'accélération des migrations (MAP)

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

migration à grande échelle

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

usine de migration

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

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

métadonnées de migration

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

modèle de migration

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

Évaluation du portefeuille de migration (MPA)

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

Évaluation de la préparation à la migration (MRA)

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

stratégie de migration

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

ML

Voir apprentissage automatique.

modernisation

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

évaluation de la préparation à la modernisation

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

applications monolithiques (monolithes)

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

MPA

Voir Évaluation du portefeuille de migration.

MQTT

Voir Message Queuing Telemetry Transport.

classification multi-classes

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

infrastructure mutable

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

O

OAC

Voir Contrôle d'accès à l'origine.

OAI

Voir l'identité d'accès à l'origine.

OCM

Voir gestion du changement organisationnel.

migration hors ligne

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

OI

Consultez la section Intégration des opérations.

OLA

Voir l'accord au niveau opérationnel.

migration en ligne

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

OPC-UA

Voir Open Process Communications - Architecture unifiée.

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

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

accord au niveau opérationnel (OLA)

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

examen de l'état de préparation opérationnelle (ORR)

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

technologie opérationnelle (OT)

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

intégration des opérations (OI)

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

journal de suivi d'organisation

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

gestion du changement organisationnel (OCM)

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

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

contrôle d'accès d'origine (OAC)

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

identité d'accès d'origine (OAI)

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

ORR

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

DE

Voir technologie opérationnelle.

VPC sortant (de sortie)

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

P

limite des autorisations

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

informations personnelles identifiables (PII)

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

PII

Voir les informations personnelles identifiables.

manuel stratégique

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

PLC

Voir contrôleur logique programmable.

PLM

Consultez la section Gestion du cycle de vie des produits.

policy

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

persistance polyglotte

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

évaluation du portefeuille

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

predicate

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

prédicat pushdown

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

contrôle préventif

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

principal

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

confidentialité dès la conception

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

zones hébergées privées

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

contrôle proactif

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

gestion du cycle de vie des produits (PLM)

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

environnement de production

Voir environnement.

contrôleur logique programmable (PLC)

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

chaînage rapide

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

pseudonymisation

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

publish/subscribe (pub/sub)

Modèle qui permet les communications asynchrones entre les microservices afin d'améliorer l'évolutivité et la réactivité. Par exemple, dans un MES basé sur des microservices, un microservice peut publier des messages d'événements sur un canal auquel d'autres microservices peuvent s'abonner. Le système peut ajouter de nouveaux microservices sans modifier le service de publication.

Q

plan de requête

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

régression du plan de requêtes

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

R

Matrice RACI

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

RAG

Voir Retrieval Augmented Generation.

rançongiciel

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

Matrice RASCI

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

RCAC

Voir contrôle d'accès aux lignes et aux colonnes.

réplica en lecture

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

réarchitecte

Voir 7 Rs.

objectif de point de récupération (RPO)

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

objectif de temps de récupération (RTO)

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

refactoriser

Voir 7 Rs.

Région

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

régression

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

réhéberger

Voir 7 Rs.

version

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

déplacer

Voir 7 Rs.

replateforme

Voir 7 Rs.

rachat

Voir 7 Rs.

résilience

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

politique basée sur les ressources

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

matrice responsable, redevable, consulté et informé (RACI)

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

contrôle réactif

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

retain

Voir 7 Rs.

se retirer

Voir 7 Rs.

Génération augmentée de récupération (RAG)

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

rotation

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

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

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

RPO

Voir l'objectif du point de récupération.

RTO

Voir l'objectif en matière de temps de rétablissement.

runbook

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

S

SAML 2.0

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

SCADA

Voir Contrôle de supervision et acquisition de données.

SCP

Voir la politique de contrôle des services.

secret

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

sécurité dès la conception

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

contrôle de sécurité

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

renforcement de la sécurité

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

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

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

automatisation des réponses de sécurité

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

chiffrement côté serveur

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

Politique de contrôle des services (SCP)

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

point de terminaison du service

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

contrat de niveau de service (SLA)

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

indicateur de niveau de service (SLI)

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

objectif de niveau de service (SLO)

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

modèle de responsabilité partagée

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

SIEM

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

point de défaillance unique (SPOF)

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

SLA

Voir le contrat de niveau de service.

SLI

Voir l'indicateur de niveau de service.

SLO

Voir l'objectif de niveau de service.

split-and-seed modèle

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

SPOF

Voir point de défaillance unique.

schéma en étoile

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

modèle de figuier étrangleur

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

sous-réseau

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

contrôle de supervision et acquisition de données (SCADA)

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

chiffrement symétrique

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

tests synthétiques

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

invite du système

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

T

tags

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

variable cible

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

liste de tâches

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

environnement de test

Voir environnement.

entraînement

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

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

passerelle de transit

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

flux de travail basé sur jonction

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

accès sécurisé

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

réglage

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

équipe de deux pizzas

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

U

incertitude

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

d'incertitude : l'incertitude épistémique est causée par des données limitées et incomplètes, alors que l'incertitude aléatoire est causée par le bruit et le caractère aléatoire inhérents aux données. Pour plus d'informations, veuillez consulter le guide Quantifying uncertainty in deep learning systems.

tâches indifférenciées

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

environnements supérieurs

Voir environnement.

V

mise à vide

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

contrôle de version

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

Appairage de VPC

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

vulnérabilités

Défaut logiciel ou matériel qui compromet la sécurité du système.

W

cache actif

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

données chaudes

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

fonction de fenêtre

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

charge de travail

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

flux de travail

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

VER

Voir écrire une fois, lire plusieurs.

WQF

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

écrire une fois, lire plusieurs (WORM)

Modèle de stockage qui écrit les données une seule fois et empêche leur suppression ou leur modification. Les utilisateurs autorisés peuvent lire les données autant de fois que nécessaire,

mais ils ne peuvent pas les modifier. Cette infrastructure de stockage de données est considérée comme immuable.

Z

exploit Zero-Day

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

vulnérabilité « jour zéro »

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

invite Zero-Shot

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

application zombie

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

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