Opérationnaliser l'IA agentique sur AWS

AWS Conseils prescriptifs

AWS Conseils prescriptifs: Opérationnaliser l'IA agentique sur AWS

Table of Contents

Opérationnaliser l'IA agentique sur AWS

Aaron Sempf, Brad Ryan, Bhargs Srivathsan et Akhil Bhaskar, Amazon Web Services

Août 2025 (historique du document)

L'IA agentic n'est pas une fonctionnalité, c'est un nouveau paradigme opérationnel. Organisations qui investissent dans une architecture rigoureuse, des cadres de confiance et des modèles de déploiement adaptés aux besoins de l'entreprise dirigeront la prochaine génération d'entreprises intelligentes et adaptatives.

L'IA agentic représente la convergence des agents logiciels autonomes et de l'IA générative. Il fusionne la prise de décision et le comportement orienté vers les objectifs des agents avec les capacités de compréhension et de génération du langage des grands modèles linguistiques (). LLMs Ces agents peuvent raisonner, agir, s'adapter et collaborer dans des environnements d'entreprise dynamiques. Pour exploiter ce potentiel, les entreprises doivent passer du déploiement de modèles à l'infrastructure d'agents.

Ce guide propose une stratégie organisationnelle pour transformer l'IA agentique issue d'expériences isolées en une infrastructure génératrice de valeur à l'échelle de l'entreprise. Il peut vous aider à intégrer des agents intelligents dans les flux de travail grâce à la gouvernance, à l'évolutivité et à l'alignement commercial.

Principaux domaines d'intérêt et recommandations

Ce guide se concentre sur les domaines fondamentaux suivants lors de l'opérationnalisation de l'IA agentique. Des recommandations organisationnelles et commerciales sont fournies pour chaque domaine d'intérêt :

- Domaine d'intervention 1 : Clarifier l'intention et le champ d'action de l'agent— Alignez les agents sur les priorités de l'entreprise et les obstacles cognitifs. Traitez les agents comme des coéquipiers numériques, et pas simplement comme des outils.
- Domaine d'intérêt 2 : Conception axée sur la composabilité et la collaboration— Adoptez des systèmes multi-agents dotés d'une architecture modulaire, de protocoles sémantiques et d'une délégation dynamique par le biais d'agents arbitres.
- Domaine d'intervention 3 : Architecte de la mutualisation et du contrôle— Créez une infrastructure évolutive adaptée aux locataires avec des services d'agents partagés, une gouvernance centralisée et un accès basé sur les rôles.

- Domaine d'intervention 4 : Instaurer la confiance grâce à l'identité, aux garde-fous et à l'observabilité— Appliquez la traçabilité, les contrôles d'exécution et l'explicabilité pour gagner la confiance des parties prenantes.
- Domaine d'intervention 5 : Gérer le cycle de vie— Établissez des pipelines d'intégration et de déploiement continus (CI/CD), une gestion rapide des versions, une télémétrie et une formation continue pour soutenir les performances et l'efficacité de l'IA agentique.
- Domaine d'intervention 6 : Aligner les modèles d'agents sur les modèles commerciaux— Monétisez les capacités des agents grâce à des modèles basés sur l'utilisation, à des indicateurs de retour sur investissement internes et à des offres commerciales.

Vous pouvez utiliser les recommandations de ce guide pour préparer votre entreprise à l'IA agentique à grande échelle. Il décrit comment les entreprises doivent se restructurer autour de l'IA agentique, notamment en mettant en place des équipes DevOps pour les agents (AgentOps), des systèmes interopérables et des stratégies de gestion du changement permettant d'étendre l'adoption. Il met l'accent sur la réflexion axée sur la décision et sur l'alignement avec le AWS Well-Architected Framework.

Public visé

Ce guide est destiné aux architectes d'entreprise, aux responsables de l' AI/ML ingénierie et aux stratèges de la transformation numérique qui conçoivent et font évoluer des systèmes agentiques, intègrent l'IA dans les principaux flux de travail commerciaux et rendent opérationnels LLMs et autonomes des agents dans les environnements de production. Pour comprendre les concepts et les recommandations de ce guide, vous devez être familiarisé avec les architectures cloud natives modernes et les systèmes distribués, les grands modèles linguistiques, les capacités des modèles de base, ainsi que les principes de gouvernance de l'IA et d'ingénierie des plateformes. DevOps

Objectifs

En mettant en œuvre les recommandations de ce guide, votre organisation peut atteindre les résultats commerciaux suivants :

- Prise de décision et exécution du flux de travail accélérées grâce à des agents autonomes et axés sur les objectifs qui réduisent les blocages humains et la charge cognitive.
- Déploiements évolutifs et rentables de capacités intelligentes au sein des unités commerciales, via des plateformes d'agents multi-locataires réutilisables.

- Renforcement de la résilience, de la confiance et de la gouvernance des systèmes d'IA, ce qui permet une adoption en toute confiance dans des environnements réglementés, critiques ou orientés client.

À propos de cette série de contenus

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

Fondements stratégiques de l'IA agentic

Les systèmes Agentic ne sont pas nouveaux. Les agents logiciels, notamment l'automatisation robotique des processus (RPA) et les moteurs de décision, existent depuis des décennies. Mais ils étaient simples et déterministes, conçus pour suivre des règles prédéfinies et une logique symbolique afin d'exécuter des tâches répétitives à faible variation. Avec l'essor de l'IA générative, la donne a changé. Les grands modèles de langage (LLMs) peuvent désormais interpréter des entrées complexes, générer des réponses de manière dynamique et synthétiser rapidement les connaissances. Vous pouvez désormais faire évoluer votre agence sans logique fragile ou codée en dur. Désormais, les agents peuvent raisonner, prendre des décisions, invoquer des outils, s'adapter au contexte et se coordonner avec d'autres agents dans tous les flux de travail. Ils peuvent agir de manière autonome pour atteindre leurs objectifs, conserver leur mémoire et réfléchir aux résultats.

Cependant, la capacité brute ne suffit pas. L'intelligence sans intégration produit de la nouveauté et non de l'impact. Pour tirer parti de la puissance LLMs, les entreprises doivent passer d'expériences isolées à des écosystèmes conçus. Les agents doivent être traités comme des services de production fonctionnant selon la même discipline que n'importe quel système d'entreprise. Cela inclut la gouvernance, l'observabilité, les modèles d'identité sécurisés et la gestion du cycle de vie. Ils doivent également se traduire par des résultats commerciaux réels, et non par un potentiel spéculatif. Ces systèmes doivent être conçus avec des limites claires pour la prise de décision et la tolérance aux pannes. Il est important d'intégrer des mécanismes de restauration automatisés, une surveillance des performances en temps réel et une gestion évolutive des ressources. Cela vous permet de gérer la nature dynamique et non déterministe des interactions avec les agents tout en maintenant des niveaux de service cohérents dans les flux de travail de l'entreprise.

À un niveau fondamental, les entreprises doivent repenser la manière dont l'intelligence est intégrée dans le tissu des opérations. Les agents doivent être conçus pour s'intégrer aux systèmes de base, respecter les politiques de l'entreprise et apporter une valeur mesurable. Ils doivent fonctionner à grande échelle, dans tous les départements, domaines et contextes utilisateurs. L'opérationnalisation de l'IA agentique est en fin de compte une question d'utilisation ; c'est la différence entre le déploiement d'une IA qui exécute des tâches isolées et le déploiement d'agents qui font évoluer votre modèle commercial.

L'IA agentic représente une nouvelle philosophie opérationnelle qui nécessite un changement fondamental dans la façon dont nous abordons les systèmes, les processus et les personnes afin d'étendre l'intelligence au sein de l'organisation. Les agents deviennent des actifs stratégiques qui amplifient les capacités humaines. En intégrant l'IA agentique à leurs opérations, les entreprises

peuvent obtenir des informations qui génèrent de la valeur commerciale, augmentent les capacités humaines et optimisent les flux de travail complexes.

Domaines d'intérêt stratégiques pour l'IA agentic

Pour passer des premiers prototypes à des systèmes de production et générateurs de valeur, les équipes ont besoin d'une stratégie cohérente alliant architecture, processus et réflexion sur le produit.

De nombreuses organisations abordent encore l'IA en privilégiant les outils ou en privilégiant les modèles. L'IA générative a amplifié l'expérimentation, mais souvent sans alignement clair sur la stratégie commerciale ou sans résultats mesurables. Sans rôle stratégique défini, les agents risquent de devenir de nouvelles expériences qui épuisent les ressources au lieu de fournir une valeur évolutive. Pour définir le rôle stratégique de l'IA agentique, les organisations doivent commencer par définir leurs priorités commerciales. Identifiez les zones de surcharge cognitive, de blocages décisionnels ou de flux de travail fragmentés où l'autonomie peut apporter un soulagement. Utilisez des énoncés de problèmes spécifiques au domaine pour définir les responsabilités des agents. Traitez les agents comme des coéquipiers numériques, et non comme des outils, capables de raisonner, de déléguer et de s'adapter.

Les sciences de la décision sont la discipline qui combine la science des données, l'analyse et la modélisation comportementale pour améliorer la prise de décision. Il doit être intégré au début du processus d'architecture des agents afin d'aligner la conception sur les résultats commerciaux. En identifiant les modèles de décision, en simulant les compromis et en quantifiant l'impact sur la valeur, les sciences de la décision peuvent vous aider à identifier les domaines dans lesquels l'autonomie agentique peut apporter le plus de valeur. Les sciences de la décision peuvent accélérer les décisions, réduire les erreurs et permettre des adaptations en temps réel. Cette base basée sur les données fonde la conception des agents sur des informations mesurables et permet une intégration plus étroite avec les technologies d'entreprise existantes, telles que les moteurs de règles, les plateformes d'analyse et les modèles prédictifs.

Pour aider à définir le rôle stratégique des agents, cette section présente les domaines d'intervention fondamentaux qui constituent l'épine dorsale de l'opérationnalisation de l'IA agentique. Chacune correspond à une tâche essentielle à accomplir du point de vue d'un responsable technique, d'un architecte ou d'un responsable de produit responsable de la façon dont les agents sont conçus et conçus. Ces domaines d'intérêt ne sont pas des étapes séquentielles. Chacun d'entre eux mérite d'être revu tout au long du cycle de vie du système afin de développer des écosystèmes d'agents résilients, évolutifs et monétisables.

Cette section contient les domaines d'intérêt suivants :

- Domaine d'intervention 1 : Clarifier l'intention et le champ d'action de l'agent

Domaine d'intervention 1 : Clarifier l'intention et le champ d'action de l'agent

Job à faire : « Aidez-moi à m'assurer que chaque agent résout un vrai problème avec des limites claires, et pas simplement une démo sympa. »

L'IA agentic ne consiste pas uniquement à renforcer les capacités. Il s'agit de résoudre le bon problème, de la bonne façon, pour obtenir le bon résultat. Cela commence par une définition parfaitement claire de l'intention de la solution d'intelligence artificielle agentic.

Stratégie

Trop souvent, les entreprises commencent par ce que le modèle peut faire (par exemple APIs, appeler, répondre à des questions ou générer des résumés) pour ensuite adapter un cas d'utilisation en fonction de celui-ci. Cela entraîne une augmentation du champ d'application, une intégration médiocre et des agents techniquement impressionnants mais inutiles du point de vue opérationnel. Commencez plutôt par définir le rôle de l'agent à l'aide de questions spécifiques telles que les suivantes :

- Quel est le résultat précis dont l'agent est responsable ?
- Au nom de qui agit-il ?
- Qui en bénéficie ?
- Où commence et où s'arrête l'autonomie de l'agent ?
- Que se passe-t-il en cas d'échec ?

Un agent bien défini a une mission claire, des responsabilités définies et des critères de réussite mesurables. Ne considérez pas l'agent comme un assistant ou un chatbot. Donnez-lui plutôt un titre

de poste. Considérez-le comme un agent de réussite client, un gestionnaire des retours de produits ou un contrôleur de conformité.

Lorsque vous impliquez des parties prenantes ou des clients, mettez l'accent sur l'évolutivité et l'adaptabilité des systèmes d'IA agentic. Ces agents évoluent avec l'entreprise et s'améliorent continuellement grâce à l'apprentissage et au feedback. Pour réduire la résistance et accélérer l'adoption, expliquez comment les outils agentic sont conçus en tenant compte de l'empathie des travailleurs. Ils fournissent de la transparence, du contrôle et des mécanismes de dérogation facultatifs qui renforcent la confiance. Plutôt que de remplacer des personnes, les agents renforcent les capacités humaines et la prise de décision, aidant ainsi les employés à rester informés et à se concentrer sur des tâches à forte valeur ajoutée.

La clé d'une mise en œuvre réussie consiste à aligner l'IA agentique sur des résultats commerciaux spécifiques et à fort impact. Encouragez les équipes et les partenaires à commencer par des projets pilotes ciblés qui résolvent les problèmes visibles. Les gains rapides génèrent un retour sur investissement (ROI) mesurable, renforcent l'adhésion interne et créent une dynamique pour une adoption plus large.

Pour guider l'adoption et la maturité, les entreprises peuvent concevoir leurs agents selon un modèle évolutif. L'autonomie, la complexité et l'impact commercial des agents augmentent progressivement. Les étapes de ce modèle sont les suivantes :

- Les agents observateurs font ressortir des informations à partir du bruit. Un agent du sentiment du marché qui suit la perception de la marque sur les canaux numériques en est un exemple.
- Les agents adjoints soutiennent la prise de décision humaine. Prenons l'exemple d'un agent de conseil en transactions qui synthétise les données sur les concurrents et les conditions du marché pour les équipes commerciales.
- Les agents autonomes agissent de manière indépendante dans des limites définies. Un agent d'allocation de ressources qui ajuste dynamiquement l'infrastructure cloud en fonction de la demande en est un exemple.
- Les agents Orchestrator coordonnent les flux de travail multi-agents. Un agent d'optimisation de la chaîne d'approvisionnement qui gère les interactions entre les agents d'inventaire, de logistique et de prévision en est un exemple.
- Les agents innovateurs génèrent de nouvelles possibilités stratégiques. Un agent d'innovation en matière de modèles commerciaux qui analyse les tendances du marché et recommande de nouvelles sources de revenus en est un exemple.

Le fait d'orienter les agents autour de ces résultats stratégiques et de ces niveaux de maturité permet de mieux se concentrer, d'accélérer l'adoption et de renforcer la confiance des parties prenantes.

Pour favoriser l'alignement dans ce domaine d'intervention Services AWS, comme Amazon Quick, vous pouvez visualiser les indicateurs de performance clés (KPIs) liés aux résultats pilotés par les agents. Vous pouvez utiliser Amazon CloudWatch pour surveiller le comportement des agents, les indicateurs de performance et l'état du système en temps quasi réel. Utilisez le feedback opérationnel pour optimiser les interactions entre les agents et l'utilisation des ressources. AWS CloudTrailpeut fournir une visibilité sur l'activité des agents et les modèles d'intégration au cours des premières phases d'expérimentation et de perfectionnement.

Valeur commerciale de la définition de l'intention et de la portée

L'adoption de l'IA agentique représente un changement crucial dans la façon dont les organisations abordent la transformation numérique et l'excellence opérationnelle. Il ne s'agit pas simplement d'automatisation. Il s'agit de permettre une autonomie intelligente qui accélère la prise de décision et la création de valeur.

Les principaux moteurs commerciaux sont les suivants :

- Avantage concurrentiel — Les premiers utilisateurs obtiennent un avantage stratégique grâce à des informations plus rapides, à un meilleur service et à des opérations adaptatives.
- Amélioration de l'expérience client : les agents offrent une assistance personnalisée et permanente en temps réel qui améliore la satisfaction et la fidélité.
- Efficacité opérationnelle — L'IA agentic réduit considérablement la charge cognitive humaine en automatisant les tâches décisionnelles complexes et répétitives. Cela permet au personnel de se concentrer sur des activités à plus forte valeur ajoutée et de réduire les coûts.

Les cas d'utilisation concrets dans tous les secteurs sont les suivants :

- Services financiers — Les agents d'intelligence artificielle pourraient fournir des conseils financiers personnalisés et détecter les fraudes.
- Soins de santé — Les agents du plan de triage et de traitement pourraient améliorer le débit clinique.
- Commerce de détail — Les agents peuvent agir comme des assistants d'achat intelligents ou optimiser les stocks en temps réel.

- Fabrication — Les agents peuvent effectuer une maintenance prédictive ou coordonner les chaînes d'approvisionnement.

Domaine d'intérêt 2 : Conception axée sur la composabilité et la collaboration

Job à faire : « Laissez-moi créer des agents comme je crée des services : modulaires et testables, afin qu'ils puissent être composés et orchestrés selon les besoins. »

De nombreux efforts en matière d'IA commencent par des projets pilotes monolithiques centrés sur le modèle. Ils sont utiles, mais ils sont difficiles à adapter à un domaine ou à un autre à des problèmes complexes. Valorisez les composés lorsque ces agents sont conçus pour interagir. En technologie, la composabilité consiste à combiner des composants modulaires pour créer une solution flexible et évolutive capable de s'adapter au changement. Sans composabilité, l'intelligence est bloquée dans des flux de travail spécifiques. En outre, la collaboration entre agents introduit des complexités en matière d'orchestration, de gestion des états et de négociation de protocoles que les équipes d'automatisation traditionnelles ne sont peut-être pas équipées pour gérer.

Stratégie

Adoptez le paradigme multi-agents. Modélisez des agents tels que des départements organisationnels : modulaires, spécialisés et interopérables. Définissez des interfaces claires, des formats de contexte partagés et des protocoles de communication standard, tels que le Model Context Protocol (MCP) ou l'Agent2Agent (A2A). Adoptez des modèles d'orchestration multi-agents, tels que la coordination en essaim, en graphe ou hiérarchique. Ces modèles aident les agents à découvrir des fonctionnalités et à demander des services les uns aux autres de manière dynamique, que ce soit dans le cadre de flux de travail parallèles, séquentiels ou basés sur le consensus, en fonction de la structure des tâches et du niveau de confiance.

Pour promouvoir une collaboration évolutive et gouvernée, utilisez un agent arbitre. Ce type d'agent est une autorité neutre qui facilite la délégation des tâches en fonction de capacités connues et de stratégies de repli. Bien qu'il ne s'agisse pas d'un contrôleur centralisé, un agent arbitre joue un rôle essentiel dans la confiance et la conformité. Il garantit que les tâches sensibles ou réglementées sont acheminées uniquement vers des agents répondant aux exigences en matière d'identité et de politique. Il agit comme un gardien pour les flux de travail liés aux politiques. Il renforce l'isolement et permet une délégation explicable. Surtout, un agent arbitre n'est pas un goulot d'étranglement ;

il coexiste avec des agents autocoordonnés qui agissent de manière horizontale. peer-to-peer Ces agents délèguent des sous-tâches, partagent le contexte et résolvent directement les dépendances.

Ce modèle hybride prend en charge à la fois l'attribution déterministe (par le biais de l'agent arbitre) et la collaboration émergente. Il allie structure et flexibilité. Dans cette architecture, les agents peuvent être classés dans les rôles spécialisés suivants :

- Agents décisionnels, tels que les responsables de l'application des politiques, les responsables de l'allocation des ressources et les évaluateurs des risques
- Agents de connaissances, tels que les agrégateurs de contexte, les outils de reconnaissance de modèles et les détecteurs d'anomalies
- Agents d'exécution, tels que les exécuteurs de tâches, les contrôleurs qualité et les responsables de l'intégration

Pour assurer une coordination efficace, les systèmes multi-agents doivent prendre en charge des protocoles d'interaction robustes pour la gestion des états, la reprise en cas de panne et la résolution des conflits. Cela favorise la stabilité et la responsabilité même lorsque les agents agissent de manière indépendante.

Établissez des règles claires pour le dimensionnement, telles que l'instanciation d'agents basée sur la charge, l'allocation des ressources adaptée au contexte et la découverte et l'enregistrement automatisés des capacités. Ces mesures aident le système à se développer de manière dynamique en réponse à la demande ou à la complexité.

Concevez les agents comme ready-to-use des modules au sein d'un substrat de messagerie distribué. Par exemple, vous pouvez utiliser Amazon EventBridge avec A2A ou MCP plutôt que des services cloisonnés. Adoptez le versionnement, les CI/CD pipelines et les modèles d'agents pour garantir la stabilité du système tout en accélérant l'adoption interne et l'évolution du cycle de vie. Encouragez la réutilisation et la standardisation du code afin de réduire les frictions liées à l'intégration et de promouvoir un écosystème résilient.

La collaboration est un multiplicateur de force. Il permet de bénéficier de l'évolutivité, de la spécialisation et de la résilience dans les environnements multi-agents. Pour soutenir cette collaboration dynamique, les entreprises doivent concevoir un plan de contrôle léger pour la coordination des agents. Ce plan de contrôle inclut les éléments suivants :

- Registres de fonctionnalités qui définissent ce que chaque agent peut faire et prennent en charge les métadonnées versionnées pour la découverte par les pairs

- Logique d'arbitrage des tâches qui utilise des agents arbitres ou superviseurs pour acheminer les tâches en fonction du contexte, de la disponibilité et de la politique
- Suivi du cycle de vie et de l'état permettant un contexte décisionnel en temps réel et des transferts sécurisés

Les plans de contrôle garantissent que les systèmes multi-agents restent extensibles, conformes aux politiques et tolérants aux pannes, sans centraliser l'autorité ni ralentir les opérations.

Cependant, les environnements multi-agents présentent également des défis opérationnels. Le maintien du contexte dans toutes les interactions avec les agents, la gestion de l'état partagé et la coordination des actions peuvent accroître la complexité et les coûts. Les coûts peuvent augmenter si vous utilisez des jetons LLMs qui consomment lors de la communication entre agents. Ces coûts doivent être mis en balance avec les avantages commerciaux combinés de l'autonomie intelligente à grande échelle.

Pour relever ces défis, considérez les plateformes agentiques qui résument les principales préoccupations, telles que les suivantes :

- Protocoles de communication et formats sémantiques standardisés
- Logique d'orchestration intégrée et routage dynamique
- Contexte partagé et gestion de la mémoire entre les agents
- Gestion des solutions de repli et dégradation progressive en cas de panne

Pour les équipes qui adoptent des stratégies multi-agents, la meilleure approche consiste à commencer modestement et à concevoir en fonction de l'échelle. Commencez par des solutions ciblées à agent unique qui résolvent de vrais problèmes. Composez ensuite progressivement ces agents dans un système coopératif dans lequel chacun peut découvrir, coordonner et déléguer en fonction d'objectifs communs et d'un contexte à l'échelle du système.

Il est important de noter que la gestion robuste des erreurs et la dégradation progressive doivent être des principes de conception fondamentaux. Les systèmes multi-agents doivent être capables de poursuivre des flux de travail partiels ou d'initier une logique de sauvegarde lorsque les agents ne sont pas disponibles ou tombent en panne. Cela favorise la fiabilité sans accouplement rigide.

Services AWS offrent des fonctionnalités robustes pour prendre en charge cette architecture à grande échelle. Amazon EventBridge et EventBridge Pipes fournissent l'épine dorsale structurée et axée sur les événements pour la messagerie multi-agents. Pour gérer le comportement modulaire,

AWS AppConfigpermet de basculer de manière sûre et dynamique entre les instances d'agent. Pour prendre en charge le contexte partagé et la gestion de la mémoire, utilisez Amazon DynamoDB pour une persistance de l'état légère et adaptée aux locataires et une récupération rapide du contexte entre les agents. Vous pouvez utiliser Amazon Simple Storage Service (Amazon S3) pour stocker des historiques d'appels structurés, des artefacts partagés ou des sorties générées par des agents. Pour les flux de travail plus complexes qui nécessitent une coordination dynamique, AWS Step Functionsvous pouvez orchestrer des processus de longue durée avec des points de contrôle et une logique de récupération des erreurs. Ensemble, ces services vous aident à créer des systèmes multi-agents composables, résilients et sémantiquement connectés qui s'adaptent aux exigences de l'entreprise.

Valeur commerciale des systèmes multi-agents

Alors que de nombreuses entreprises commencent leur parcours vers l'IA avec des solutions à agent unique, le plein potentiel de l'IA agentique est exploité grâce à des systèmes multi-agents évolutifs. Ces systèmes sont essentiels pour résoudre des problèmes complexes et distribués et créer des écosystèmes d'IA robustes et flexibles qui évoluent en fonction des besoins de l'entreprise.

Les principaux avantages commerciaux des systèmes multi-agents sont les suivants :

- Évolutivité — Les tâches et les charges de travail peuvent être réparties entre des agents spécialisés afin d'augmenter la capacité et les performances.
- Flexibilité : les agents peuvent être ajoutés, remplacés ou modifiés avec un minimum de perturbations, ce qui permet une certaine agilité dans les environnements dynamiques.
- Résilience — La stabilité du système est préservée même en cas de défaillance d'agents individuels, grâce à des rôles redondants et à un basculement intelligent.
- Spécialisation — Les agents spécialement conçus exécutent les tâches avec une efficacité et une précision accrues.
- Rentabilité — Les composants réutilisables des agents accélèrent le développement et réduisent le coût du déploiement de nouvelles fonctionnalités.

Bien que les systèmes multi-agents nécessitent une planification initiale plus poussée, ils offrent agilité, rapidité et capacité d'innovation à long terme. Les entreprises qui investissent dans des architectures flexibles de collaboration avec des agents sont bien placées pour déployer rapidement de nouvelles capacités d'IA, s'adapter à l'évolution des demandes et occuper une position de leader dans un environnement concurrentiel de plus en plus axé sur les agents.

Domaine d'intervention 3 : Architecte de la mutualisation et du contrôle

Job à accomplir : « Aidez-moi à adapter l'utilisation des agents à plusieurs clients sans perdre le contrôle, la responsabilité ou la visibilité. »

Les premiers prototypes permettent de prouver la valeur de manière isolée, mais la plupart des entreprises doivent prendre en charge simultanément plusieurs clients, départements ou flux de travail. Cela signifie que chaque agent doit opérer dans le cadre de politiques, de données et de limites d'identité clairement définies. Sans la mutualisation, les opérations deviennent fragiles et coûteuses, et la gouvernance devient disparate.

Stratégie

Suivez les principes des architectures SaaS (Software as a Service). Par exemple, conception pour isoler les locataires, appliquer les politiques et contrôler les ressources. Architectez des agents et des plateformes d'orchestration avec une mémoire, une configuration et une identité adaptées aux locataires. Pour faire respecter les limites, utilisez le balisage, le contrôle d'accès basé sur les rôles (RBAC) et le périmètre de gestion des identités et des accès.

Adoptez une couche d'observabilité unifiée dans laquelle la télémétrie des agents est agrégée par contexte de locataire. Mettez en œuvre des moteurs de politiques centralisés et un basculement des capacités basé sur la configuration pour appliquer des règles de comportement dynamiques.

Développez le déploiement d'agents en tant que service. Permettez aux équipes internes ou aux clients d'utiliser les capacités des agents de manière évolutive et gouvernée APIs. AWS fournit une base solide pour ces modèles. Vous pouvez utiliser Amazon Cognito pour gérer l'identité des utilisateurs et des locataires, AWS Organizationsainsi que les politiques de contrôle des services (SCPs) pour la gouvernance entre comptes et AWS Resource Access Manager (AWS RAM) pour partager des fonctionnalités de partage en toute sécurité. En outre, il AWS AppConfigpeut gérer dynamiquement le comportement des agents par locataire ou par environnement. Ces services aident à faire respecter les limites et les politiques tout en soutenant l'infrastructure partagée.

Cette transition du déploiement statique au provisionnement dynamique transforme l'IA agentic en une plateforme à l'échelle de l'entreprise.

Valeur commerciale des plateformes d'agents multi-locataires

La mutualisation est bien plus qu'une simple commodité architecturale, c'est un accélérateur commercial. Alors que les agents intelligents se multiplient entre les départements et les équipes, les entreprises doivent soutenir la croissance sans dupliquer l'infrastructure ni fragmenter la gouvernance.

Les principaux avantages commerciaux des systèmes multi-locataires sont les suivants :

- Évolutivité — Une plateforme d'agents multi-locataires permet aux équipes internes, aux unités commerciales ou aux clients d'intégrer les fonctionnalités d'IA plus rapidement sans avoir besoin d'environnements sur mesure.
- Rentabilité — L'infrastructure partagée minimise les déploiements redondants, consolide les coûts d'exploitation et simplifie la maintenance dans tous les environnements.
- Gouvernance et réduction des risques — Les contrôles stratégiques centralisés, les modèles d'identité et l'observabilité aident les agents à fonctionner de manière plus sécurisée et conforme, pour tous les locataires.
- Réutilisabilité des services — Pour promouvoir la réutilisation et réduire les doublons, des agents sensibles aux locataires peuvent être proposés en tant que services internes, par exemple pour l'enrichissement, la conformité ou le résumé.

Les exemples d'utilisation des systèmes à locataires multiples sont les suivants :

- Un agent de conformité déployé dans les filiales adapte sa logique aux réglementations locales grâce à une configuration spécifique au locataire. Il n'est donc plus nécessaire de créer des agents distincts pour chaque région.
- Un agent interne d'automatisation des flux de travail dessert plusieurs départements avec des limites de données et des autorisations différentes. Il maintient l'isolement tout en accélérant l'exécution des tâches.

En concevant les agents comme des multi-tenant-aware services, les entreprises évitent les frais généraux liés à des initiatives d'IA cloisonnées. Ils favorisent plutôt une plateforme de renseignement unifiée. Cette architecture permet un déploiement évolutif, une cohérence opérationnelle et un meilleur retour sur investissement. Cela permet également d'étendre plus facilement l'adoption de l'IA au sein de l'entreprise.

Domaine d'intervention 4 : Instaurer la confiance grâce à l'identité, aux garde-fous et à l'observabilité

Job à faire : « Donnez-moi l'assurance que les agents agiront de manière sûre et prévisible, en particulier lorsque personne ne les regarde. »

Les agents autonomes remettent en question les modèles de contrôle traditionnels. Leur capacité à raisonner et à agir de manière indépendante présente des risques s'ils ne sont pas correctement gérés. En l'absence de contraintes claires en matière de propriété, d'auditabilité ou de politique, ils peuvent s'écarter du comportement prévu. L'établissement de la confiance organisationnelle ne se limite pas à la fiabilité technique. Cela exige de l'explicabilité, de la responsabilité et de la cohérence.

Stratégie

Construisez un système de contrôle axé sur l'identité comme fondement d'une autonomie fiable. Chaque agent doit fonctionner avec une identité vérifiable, des autorisations limitées et un historique d'exécution traçable. Les agents doivent être intégrés dans un cadre de confiance zéro qui inclut la liaison entre locataires, l'héritage des accès contextuels et l'application du temps d'exécution par le biais de garde-fous et de moteurs de politiques. Cela vous permet d'auditer, d'annuler ou de restreindre les actions des agents en fonction des règles organisationnelles et du niveau de risque.

Intégrez le renforcement de la confiance au moment de l'exécution grâce à des garde-fous intelligents. Cela inclut le contrôle des taux et la régulation basés sur les modèles comportementaux ou les conditions de charge de travail, les limites des ressources appliquées parallèlement à l'auto-scalage et la notation des décisions pour évaluer les risques. Créez des déclencheurs pour engager les human-in-the-loop flux de travail lorsque les seuils sont dépassés.

Chaque agent doit également être transparent et explicable. Intégrez la télémétrie structurée par le biais de la journalisation, des traces et des résumés de raisonnement pour exposer la logique décisionnelle. Support des pistes décisionnelles et du suivi de l'impact. Cela vous permet de relier les actions des agents aux indicateurs ou aux résultats clés. Mettez en œuvre des mécanismes de détection des dérives qui surveillent les écarts par rapport aux comportements ou aux politiques attendus.

Introduisez des agents réfléchissants qui observent en permanence le comportement des agents et les modèles du système. Ils doivent signaler les anomalies ou les incohérences en temps réel. Ces agents contribuent aux boucles de rétroaction de gouvernance qui peuvent déclencher la revalidation, l'adaptation ou la mise hors service des fonctionnalités.

Établissez des conseils de gouvernance qui examinent les politiques relatives aux agents, approuvent les modifications des capacités et supervisent les protocoles de réponse aux incidents. La confiance doit être gagnée, mesurée et continuellement renforcée.

AWS fournit une base solide pour la mise en œuvre de ce cadre de confiance :

- Gestion des identités et des accès AWS (IAM) applique l'exécution basée sur les rôles et les limites d'autorisation
- Amazon CloudWatch et AWS X-Raysoutiennent une visibilité et une traçabilité complètes.
- Amazon GuardDuty et AWS Configdétectez les anomalies de sécurité ou les dérives politiques.

Ensemble, ces services permettent le renforcement des identités, la sécurité des environnements d'exécution et une gouvernance basée sur la confiance à grande échelle. Ils peuvent contribuer à rendre les systèmes autonomes à la fois puissants et fiables.

La valeur commerciale d'une autonomie fiable

À mesure que les agents gagnent en autonomie, la confiance devient un moteur essentiel pour l'adoption, la gouvernance et les performances opérationnelles de l'entreprise. L'établissement d'une base d'identité, d'observabilité et de garde-fous aide les entreprises à étendre l'IA agentique à des domaines sensibles, sans pour autant sacrifier la gouvernance ou le contrôle.

Les principaux moteurs commerciaux sont les suivants :

- Assurance de la gouvernance — Des modèles d'identité, des pistes d'audit et des limites d'autorisation solides réduisent les risques de conformité et favorisent l'alignement réglementaire.
- Continuité opérationnelle : les garde-fous d'exécution et la détection des anomalies aident à prévenir les comportements imprévus et à favoriser l'autoréparation en cas de défaillance ponctuelle.
- Confiance des parties prenantes — L'explicabilité des décisions et la télémétrie renforcent la confiance des parties prenantes internes, des gestionnaires des risques et des auditeurs externes.
- Résilience aux incidents : l'observabilité intégrée accélère l'analyse des causes profondes et le temps de réponse en cas de problème.

Les exemples de cas d’utilisation incluent :

- Dans les services financiers, les agents de détection des fraudes doivent exposer leur raisonnement, enregistrer chaque action avec une identité traçable et opérer dans le cadre de rôles IAM étroitement définis.
- Dans le secteur de la santé, les agents de triage autonomes doivent appliquer des contrôles de sécurité en cours d'exécution, passer à un examen humain lorsque les seuils sont atteints et fournir des journaux complets à des fins de surveillance clinique.

En intégrant des mécanismes de confiance dans le cycle de vie des agents, les entreprises peuvent permettre à leurs systèmes de fonctionner de manière autonome et responsable. Cette base réduit les risques et permet aux agents d'agir au nom de l'entreprise avec transparence et intégrité.

En fin de compte, l'autonomie fiable accélère l'adoption en donnant aux utilisateurs et aux dirigeants la confiance nécessaire pour faire évoluer les agents intelligents dans l'ensemble des opérations principales.

Domaine d'intervention 5 : Gérer le cycle de vie

Job à faire : « Faire en sorte que mon équipe puisse améliorer ses agents au fil du temps, sans chaos ni héroïsme. »

Contrairement aux applications traditionnelles qui ne sont façonnées que par le code, le comportement des agents est également façonné par les instructions, la mémoire, les outils et le contexte de formation. Ces facteurs évoluent au fil du temps. La dérive nuit à la fiabilité, augmente les coûts et rend le débogage quasiment impossible. Sans contrôle du cycle de vie, les agents cessent de fournir de la valeur et commencent à accumuler des risques.

Stratégie

Établissez une pratique DevOps pour les agents (AgentOps). Intégrez des CI/CD pipelines adaptés aux agents. Utilisez ces pipelines pour tester les sorties rapides, valider les intégrations d'outils et établir le profil coût-performance. Conservez l'historique des versions des invites, des politiques et des interactions entre les modèles.

Utilisez les boucles de feedback issues des données d'observabilité pour initier le recyclage, le réglage rapide ou le retrait des agents. Intégrer des mécanismes de réflexion à l'échelle du système, tels qu'un registre d'amélioration, pour institutionnaliser l'apprentissage.

Créez un tableau de bord de télémétrie des performances qui indique la précision des décisions, la latence, le coût et la fiabilité. Pour rationaliser et accélérer la gestion du cycle de vie à l'aide de

l' AWS infrastructure, les équipes peuvent utiliser des boîtes à outils pour agents. Le SDK Strands Agents en est un exemple. Il fournit des outils structurés pour une gestion rapide des versions, un enregistrement des outils et une intégration CI/CD avec Services AWS, par exemple, et. AWS CodePipelineAWS Cloud Development Kit (AWS CDK)AWS Lambda Utilisez également Amazon S3 et Amazon Elastic File System (Amazon EFS) pour stocker les artefacts des agents et les données de formation. Utilisez-le AWS Step Functionspour automatiser les flux de travail complexes de reconversion ou de validation. Vous pouvez utiliser Amazon SageMaker AI lorsque les agents ont besoin d'un réglage personnalisé du modèle ou d'un ajustement précis des flux de travail au-delà de l'orchestration du LLM. La discipline du cycle de vie transforme les agents issus des expériences en actifs durables et évolutifs.

Au fil du temps, ce système de cycle de vie constitue l'épine dorsale de l'innovation. Il vous aide à recomposer, à réentraîner et à redéployer les fonctionnalités avec agilité. Cela transforme la couche d'agents en un système vivant, capable d'évoluer en réponse à la fois aux commentaires et aux opportunités.

Valeur commerciale de la gestion du cycle de vie

La gestion efficace du cycle de vie est un facteur clé des performances des agents et de la rentabilité. Cela garantit que les agents intelligents continuent de fournir des résultats précis, fiables et conformes à la valeur au fur et à mesure de leur évolution. Les agents ne conservent pas leur valeur par défaut. Ils doivent évoluer en fonction de l'évolution des exigences commerciales, des flux de travail et des environnements de données. Une AgentOps équipe disciplinée aide les agents à rester précis, efficaces et alignés sur les objectifs de l'entreprise au fil du temps.

Les principaux moteurs commerciaux sont les suivants :

- Cohérence des performances — Les tests continus, la validation rapide et le recyclage aident les agents à maintenir la qualité des décisions dans des conditions et des ensembles de données changeants.
- Optimisation des coûts — Le profilage basé sur la télémétrie permet d'identifier les outils inefficaces, les demandes trop importantes ou les exécutions inutiles. Vous pouvez ensuite effectuer des réglages pour réduire les coûts d'exploitation.
- Itération accélérée — L'automatisation du cycle de vie CI/CD accélère les cycles de développement, aidant les équipes à expérimenter, déployer et améliorer les agents en toute confiance.

- Réduction des risques — La gestion rapide des versions, la prise en charge des annulations et les mécanismes d'évaluation structurés aident à prévenir les régressions et à assurer une gestion des modifications sûre et fiable.

Les exemples de cas d'utilisation incluent les suivants :

- Un agent du support client est surveillé en fonction de la latence, du coût du modèle et des commentaires des utilisateurs. L'observabilité révèle une hausse des coûts, ce qui entraîne un réajustement des instructions intégrées et de la logique du modèle de repli.
- Un agent de synthèse des contrats est mis à jour en fonction des commentaires des équipes juridiques. Les instructions versionnées sont testées dans des environnements sandbox avant leur sortie de production, afin de garantir la sécurité et la qualité.

Grâce à une gestion structurée du cycle de vie, les entreprises passent de la maintenance réactive à une amélioration continue et proactive. Les agents deviennent des actifs numériques adaptatifs mesurés, affinés et revalidés par rapport aux objectifs commerciaux. Cette pratique transforme les écosystèmes d'agents en systèmes hautement performants, sensibles aux coûts et résilients qui offrent une valeur durable tout en suivant le rythme du changement.

Domaine d'intervention 6 : Aligner les modèles d'agents sur les modèles commerciaux

Job à faire : « Montrez-moi l'impact, afin que je puisse justifier la poursuite des investissements. »

Même les agents techniquement compétents deviennent des passifs s'ils ne sont pas liés aux résultats commerciaux. Les agents doivent être au service de l'efficacité, de la monétisation ou de la différenciation stratégique. Pourtant, la plupart des entreprises ont du mal à définir la place des agents dans les modèles de tarification, d'emballage ou d'utilisation. Sans un alignement clair sur la valeur commerciale, il est difficile de justifier la mise à l'échelle ou même le maintien de l'investissement.

Stratégie

Adoptez des pratiques de gestion des produits. Traitez les agents comme des services monétisables avec un retour sur investissement mesurable. Définissez des stratégies de tarification en fonction des décisions, des sessions ou des résultats. Ensuite, regroupez les capacités des agents dans des offres à plusieurs niveaux adaptées aux segments de clientèle ou aux unités commerciales internes.

Pour promouvoir le développement durable, les entreprises doivent saisir à la fois la valeur directe et les multiplicateurs de croissance grâce au déploiement d'agents. Envisagez d'utiliser les indicateurs de retour sur investissement suivants pour mesurer la valeur immédiate :

- Coût par décision — Comparez les coûts de traitement des agents par rapport aux équivalents humains.
- Compression temporelle : quantifiez la valeur des cycles accélérés, tels que l'accélération des ventes ou des approbations.
- Réduction des erreurs : mesurez les économies réalisées grâce à l'amélioration de la précision, de la cohérence et de la conformité.

Au-delà de ces gains immédiats, les agents peuvent débloquer les opportunités de croissance à long terme suivantes :

- Cumul des capacités : combinez les services des agents pour créer des solutions verticales spécifiques au domaine.
- Effets de réseau — Augmentez la valeur grâce à des écosystèmes multi-agents dans lesquels la coordination joue un rôle utile.
- Extension du marché — Générez de nouvelles sources de revenus grâce à des services consommables externes et assistés par des agents.

Créez des boucles de feedback à partir de statistiques commerciales (telles que les économies de coûts, l'augmentation des conversions, time-to-resolution etc.) pour favoriser l'évolution continue des agents. Analysez la télémétrie d'utilisation et les scores de satisfaction des utilisateurs pour affiner l'alignement des valeurs et les priorités de votre feuille de route. En liant directement les capacités des agents aux modèles commerciaux, les entreprises se positionnent de manière à obtenir une valeur durable et cumulable, et pas seulement des résultats techniques.

Les éléments suivants Services AWS soutiennent cet alignement en fournissant des cadres de suivi et de monétisation robustes :

- AWS Cost Exploreret Amazon CloudWatch fournissent des informations sur les coûts par agent et l'efficacité opérationnelle.
- Amazon API Gateway permet un accès limité, une limitation de débit et une tarification échelonnée pour les points de terminaison des agents.

- AWS Marketplacefournit un canal aux agents de publication et aux solutions agentiques en tant que produits commerciaux.

Ces services vous aident à transformer les fonctionnalités des agents en offres numériques évolutives et axées sur la valeur qui s'alignent sur les stratégies de croissance et de monétisation de l'entreprise.

Livraison logicielle évolutive pour l'IA agentic

La fourniture de logiciels modernes repose sur une hypothèse simple : vous contrôlez les systèmes que vous expédiez. Vous définissez les exigences, rédigez la logique, testez par rapport aux résultats attendus et déployez des services prévisibles. Même l'agilité et les DevOps approches reposent toujours sur le principe selon lequel chaque sprint apporte quelque chose de déterministe, de vérifiable et largement supervisé par un humain.

L'IA agentic bouleverse cette base. Les systèmes agentic interprètent, raisonnent et adaptent plutôt que de suivre des scripts. Leur comportement dépend du code que vous écrivez, du contexte dans lequel ils opèrent, des entrées qui leur sont fournies, des outils auxquels ils peuvent accéder et des objectifs qui leur sont assignés. Bref, ils ne suivent pas les ordres ; ils recherchent des résultats.

La livraison est donc moins une question de contrôle qu'une question d'alignement. Plutôt que de fournir des instructions, vous devez façonner son comportement. Cela signifie que le cycle de vie de développement logiciel (SDLC) traditionnel n'est plus adapté car il a été conçu pour des systèmes basés sur la logique et contrôlés par l'homme.

Cette section contient les rubriques suivantes :

- Zones d'intention pour l'IA agentic
- Évolution du cycle de vie de livraison pour l'IA agentic
- Préparer les équipes à l'IA agentic

Zones d'intention pour l'IA agentic

Au lieu d'étapes rigides, telles que la définition, la construction, le test et la publication, nous avons besoin d'un modèle qui intègre l'autonomie, l'incertitude et l'émergence. Vous utilisez plutôt des zones d'intention. Une zone d'intention définit un espace délimité dans lequel un agent peut opérer de manière autonome, dans le cadre de contraintes. L'objectif est de passer de la microgestion de chaque tâche à la conception d'environnements dans lesquels les agents peuvent agir, apprendre et collaborer en toute sécurité. Vous spécifiez le quoi (le résultat souhaité), le pourquoi (l'intention) et les garde-fous (les contraintes, les politiques et les limites de confiance). Compte tenu de ces limites et de ces informations, l'agent détermine comment.

Au lieu d'une chaîne de montage, imaginez l'environnement comme un espace aérien. Vous contrôlez qui peut entrer, ce qu'ils peuvent faire et où ils peuvent aller. Mais une fois à l'intérieur, ils

sont libres de naviguer selon leurs besoins. C'est ainsi que les systèmes agentiques évoluent sans chaos.

Il ne s'agit pas simplement d'un changement philosophique ; c'est un changement pratique. Les résultats non déterministes des systèmes basés sur des agents ne peuvent pas être entièrement testés par le biais de tests unitaires. Il ne peut pas être versionné comme les binaires statiques. Les agents évoluent au fil du temps, s'adaptent aux nouvelles données et interagissent avec d'autres systèmes de manière imprévisible. Essayer de les fournir à l'aide de modèles traditionnels conduit à des architectures fragiles et non évolutives. Au pire, cela conduit à une fausse confiance dans des systèmes que vous ne pouvez pas réellement gouverner.

Lorsque les équipes adoptent la livraison basée sur l'intention, elles bénéficient de deux avantages :

- Contrôlez là où cela compte le plus : ils définissent des limites plutôt que des résultats.
- Évolutivité grâce à la délégation : ils permettent aux agents de gérer une complexité que les humains ne peuvent pas coder en dur.

C'est ainsi que vous passez de prototypes isolés à de véritables systèmes agentiques de production capables de générer de la valeur de manière répétée et fiable.

Évolution du cycle de vie de livraison pour l'IA agentic

Pour favoriser un comportement intelligent et adaptatif, le SDLC doit être redéfini pour passer d'un contrôle déterministe à une intention adaptative. Voici les modifications nécessaires pour faire évoluer le SDLC traditionnel pour l'IA agentique :

- La planification devient un design d'intention. Les équipes définissent les objectifs, les contraintes et les comportements attendus des agents. Les politiques et les critères de réussite sont définis en termes d'alignement et non de logique.
- L'architecture devient un échafaudage. Les équipes se concentrent sur la définition des rôles, des interfaces, des garde-fous, des mécanismes de repli et de l'observabilité plutôt que sur l'élaboration de scripts pour chaque processus décisionnel.
- Les tests deviennent une évaluation comportementale. Plutôt que d'affirmer des résultats spécifiques, les équipes vérifient si les agents respectent les limites acceptables et répondent à leurs intentions en fonction de divers intrants.

- Le déploiement devient une orchestration continue. Les systèmes Agentic sont déployés avec des contrôles d'exécution, une surveillance en direct et des canaux de feedback qui permettent un réglage en temps réel.
- L'itération devient feedback et adaptation. Au lieu des cycles traditionnels de modification de code, les équipes observent comment les agents évoluent, où ils réussissent ou quand ils dérivent. Au besoin, les équipes interviennent en actualisant les contraintes, en se reformant et en ajoutant ou en modifiant des mécanismes de contrôle.

Les pratiques existantes qui mettent l'accent sur l'itération, l'expérimentation et le feedback rapide sont à mi-chemin. Le passage aux systèmes agentiques n'est pas un rejet des principes agiles. En fait, il s'agit d'une évolution naturelle de celles-ci. La pensée agile met l'accent sur l'adaptabilité, le feedback et les solutions de travail plutôt que sur des plans rigides. Cela correspond parfaitement à la nature des systèmes agentiques, qui apprennent, s'adaptent et réagissent au contexte en temps réel. Si vous utilisez déjà des cycles courts, que vous validez rapidement des hypothèses et que vous gérez l'incertitude grâce à une livraison continue, vous êtes bien équipé pour mener cette transition.

Mais il existe des différences fondamentales. L'approche agile traditionnelle suppose que le produit livré est déterministe. Cela suppose qu'une fois construit, l'objet se comportera de manière cohérente et prévisible, avec des résultats reproductibles pour les mêmes entrées. Cette répétabilité vous permet de déboguer, de tester et d'itérer en toute confiance. Les systèmes agentiques bouleversent ce modèle. Ils sont probabilistes, sensibles au contexte et capables d'évoluer de manière indépendante. Cela signifie que certaines pratiques agiles deviennent moins utiles, telles que le suivi de la vélocité basé sur l'achèvement des histoires, des critères d'acceptation stricts ou une planification déterministe des sprints.

Les aspects suivants du SDLC traditionnel s'appliquent à l'IA agentic :

- Développement et livraison itératifs
- Les commentaires des clients comme signal principal
- Collaboration interfonctionnelle
- Intégration et déploiement continus

Les aspects suivants du SDLC traditionnel doivent évoluer pour l'IA agentic :

- Redéfinissez le fait comme étant aligné sur l'intention. Concentrez-vous sur la question de savoir si le comportement de l'agent répond à son objectif dans le cadre des contraintes définies.

- Passez des critères d'acceptation aux barrières comportementales.
- Élargissez la définition de « fait » pour inclure la préparation à l'exécution, qui inclut les mécanismes d'observabilité, d'explicabilité et de feedback qui favorisent l'apprentissage continu et la confiance.
- Privilégiez les boucles de feedback en temps réel et le suivi des comportements par rapport à la planification initiale

La bonne nouvelle, c'est que vous n'avez pas besoin de jeter le playbook SDLC. Il vous suffit de le faire évoluer, de la gestion du code à l'élaboration du comportement. Dans les systèmes agentiques, le succès ne dépend pas seulement de l'exécution du logiciel, mais de son comportement.

Préparer les équipes à l'IA agentic

Le génie logiciel ne va pas disparaître. Elle évolue. Le travail passe de l'écriture de fonctions à la mise en place de cadres et de mécanismes de contrôle pour un comportement intelligent. Dans le monde de l'IA agentique, le plus difficile n'est plus de construire, mais de gérer l'émergence. Pour la plupart des équipes d'ingénierie, cette évolution ressemble à un changement de mentalité plutôt qu'à un saut technique. Au lieu de demander « Que fera le système ? » la question devient : « Qu'est-ce que nous lui avons donné les moyens de poursuivre, et comment saurons-nous s'il maintient le cap ? »

Pour les équipes d'ingénierie, l'évolution vers l'IA des agents nécessite les changements suivants :

- Un changement culturel — Les équipes doivent se familiariser avec l'incertitude et l'autonomie dans des systèmes qu'elles ne contrôlent pas totalement.
- Nouveaux rôles — Les concepteurs d'intentions, les testeurs comportementaux et les ingénieurs d'observabilité jouent un rôle essentiel dans la prestation de services.
- Langage partagé : les équipes ont besoin d'une compréhension claire et partagée des objectifs, des garde-fous et des signaux de réussite, comme elles avaient auparavant besoin de spécifications et de scénarios de test.

À mesure que l'IA générative arrivera à maturité, nous verrons de plus en plus de systèmes agentiques interagir avec les clients, les produits et les opérations. Les organisations qui réussiront ne seront pas celles qui auront les meilleurs modèles. Ce seront eux qui pourront intégrer les agents dans des flux de travail réels en toute confiance, contrôle et rapidité. Cela signifie que les modèles de prestation et les équipes d'ingénierie doivent évoluer ensemble. Les zones d'intention vous donnent

l'abstraction nécessaire pour le faire. Ils vous aident à rendre votre autonomie opérationnelle sans renoncer à la responsabilité. Ils offrent également un cadre partagé entre les équipes pour aider à gouverner les systèmes qui ne peuvent pas être codés en dur.

Pour plus d'informations sur la préparation des équipes à l'IA agentique, consultez la section Préparer l'entreprise à l'IA agentique à grande échelle de ce guide.

Préparer l'entreprise à l'IA agentique à grande échelle

Au fur et à mesure que les domaines d'intérêt décrits dans ce guide convergent, l'IA agentique passe de fonctions isolées à une couche d'intelligence unifiée qui peut être comprise comme une plate-forme de capacités. Cette plateforme ne se contente pas d'exécuter des tâches. Il évolue, s'adapte et assure la coordination entre les domaines. Les agents deviennent des services modulaires, réutilisables et détectables qui accélèrent l'innovation, réduisent la charge cognitive et génèrent des résultats mesurables au sein de l'entreprise. Cette vue de plateforme ouvre la voie à une intelligence évolutive intégrée à l'ensemble du modèle d'exploitation.

L'opérationnalisation de l'IA agentique ne se limite pas au déploiement d'agents intelligents. Cela exige une transformation fondamentale de la façon dont l'entreprise organise les équipes, conçoit les processus et gouverne la technologie. Tout comme le passage au cloud ou à des modèles d'exploitation DevOps redéfinis, l'intelligence artificielle agentic ouvre une nouvelle ère d'automatisation des décisions, d'apprentissage continu et de coordination autonome. Le succès dépend de l'alignement des systèmes, des personnes et des processus autour de cette nouvelle philosophie opérationnelle.

Cette section contient les rubriques suivantes :

- Harmoniser les équipes et les modèles de propriété
- Gestion du changement et préparation organisationnelle
- Architecture axée sur l'interopérabilité et la collaboration
- Intégrer la gouvernance dans un tissu agentique
- Adopter un état d'esprit opérationnel axé sur les décisions
- Évoluer en fonction du but et de l'intention

Harmoniser les équipes et les modèles de propriété

La première étape vers la maturité est l'alignement interfonctionnel. Les entreprises doivent mettre en place des AgentOps équipes comprenant des AI/ML praticiens et des spécialistes du domaine, tels que des architectes de systèmes distribués, des ingénieurs logiciels, des responsables de produits, des responsables de la conformité et des architectes de plateformes. Ces équipes sont conjointement responsables de l'ensemble du cycle de vie d'un agent, de la conception au déploiement, en passant par la formation continue et le suivi.

Le provisionnement et la publication des agents doivent suivre les pratiques natives du cloud, telles que l'utilisation de AWS Cloud Development Kit (AWS CDK)et AWS CodePipelinepour l'infrastructure sous forme de code et le déploiement automatisé. Cette structure favorise le partage des responsabilités et accélère l'itération. Tout comme il DevOps unifie le développement et les opérations, il AgentOps associe l'intelligence à la gouvernance et à l'exécution.

Pour être efficaces, ces équipes ont également besoin d'un langage commun. Les parties prenantes de l'entreprise doivent comprendre ce que sont les agents, leur mode de fonctionnement et les résultats qu'ils obtiennent. La formation et l'habilitation interne sont essentielles. En démystifiant les agents et en intégrant ce modèle mental dans les conversations quotidiennes, les organisations favorisent une participation plus large et une innovation mieux harmonisée.

Pour accélérer le développement et l'intégration des agents qui les utilisent Services AWS, les équipes peuvent adopter des frameworks tels que le SDK Strands Agents, qui propose des outils basés sur une interface de ligne de commande pour l'échafaudage, la configuration et le packaging des agents. Strands Agents est conçu pour fonctionner parfaitement avec AWS des infrastructures telles qu'Amazon Bedrock AWS Lambda, Amazon EventBridge, le AWS CDK, et AWS CodePipeline. Il permet un prototypage et un déploiement rapides tout en respectant les normes de production.

Mais la structure et l'outillage ne suffisent pas à eux seuls. Le développement de l'IA agentique nécessite une volonté délibérée en matière de culture, d'éducation et de leadership afin de garantir que l'adoption prenne racine dans l'ensemble de l'organisation.

Gestion du changement et préparation organisationnelle

Pour réussir à faire évoluer l'IA agentic, il ne suffit pas de déployer une infrastructure ou des agents intelligents. Cela exige une approche structurée du changement organisationnel. Cela inclut la préparation culturelle, le développement des compétences, les boucles de feedback basées sur des indicateurs et l'alignement de la direction pour s'assurer que l'adoption est à la fois intentionnelle et durable.

Favoriser l'évolution culturelle

- Positionnez les agents comme des coéquipiers, et non comme des remplaçants, afin de réduire la résistance et de renforcer la confiance.
- Communiquez de manière transparente sur les capacités et les limites des agents afin de définir des attentes réalistes.

- Établissez des protocoles de transfert clairs indiquant dans quels cas les agents doivent transmettre les décisions à une autorité supérieure ou déléguer certaines parties du processus à un collaborateur humain.

Mettre en place un cadre de développement des compétences

- Offrez une formation basée sur les rôles adaptée aux ingénieurs, aux chefs de produit, aux responsables de domaine et aux responsables de la conformité.
- Créez des centres d'excellence pour partager les meilleures pratiques, les modèles d'outillage et les actifs réutilisables.
- Associez des spécialistes de l'IA à des experts du domaine par le biais de programmes de mentorat pour combler les lacunes en matière de connaissances.

Définissez des métriques et des boucles de feedback

- Ancrez la valeur technique et commerciale KPIs à la valeur stratégique pour évaluer l'impact. Les exemples de valeur incluent la latence des décisions, la précision de la résolution et les économies de coûts.
- Recueillez systématiquement et en continu les commentaires des utilisateurs sur les points de friction des surfaces et les défis d'adoption.
- Réalisez régulièrement des rétrospectives pour évaluer les performances des agents, les tendances d'utilisation et les opportunités d'amélioration.

Aligner le leadership depuis le sommet

- Obtenez le parrainage de la direction en liant les initiatives des agents aux résultats stratégiques et au retour sur investissement.
- Formez des comités de gouvernance interfonctionnels comprenant à la fois des dirigeants techniques et commerciaux.
- Personnalisez les stratégies de communication pour plus de clarté et d'engagement à tous les niveaux de l'organisation.

Cette approche systématique de la gestion du changement garantit que la mise en œuvre de la technologie correspond à la maturité organisationnelle. Il jette les bases de la confiance, de l'adoption et de la valeur commerciale à long terme.

Architecture axée sur l'interopérabilité et la collaboration

Les déploiements d'agents isolés offrent des avantages locaux. Mais la valeur d'entreprise émerge lorsque les agents peuvent découvrir, invoquer et collaborer les uns avec les autres de manière dynamique. Cela implique de définir des normes pour l'enregistrement des agents, l'authentification et l'échange de capacités. Sur le plan architectural, cela reflète le passage des monolithes aux microservices, qui sont des unités composables, réutilisables et faiblement couplées qui résolvent ensemble des problèmes complexes.

Les protocoles émergents, tels que l'A2A et le MCP, sont fondamentaux. Ils permettent l'interopérabilité sémantique entre les agents, les outils et les systèmes de mémoire. A2A prend en charge l'interaction entre pairs, ce qui permet aux agents de négocier la propriété des tâches, de partager le contexte et de coordonner les flux de travail. MCP complète cela en proposant des schémas partagés pour l'échange de données contextuelles entre les agents et leurs environnements. Il normalise la manière dont les fonctions sont invoquées, APIs accessibles et les états sont conservés. Ensemble, ces protocoles favorisent l'extensibilité, la cohérence et la maintenabilité à long terme dans l'ensemble de l'écosystème des agents.

La gouvernance demeure essentielle. Les couches de contrôle, telles que les agents arbitres, permettent une délégation adaptée aux politiques sans introduire de goulets d'étranglement centralisés. Ces agents agissent en tant que courtiers fiduciaires. Ils imposent des limites tout en laissant les autres agents s'auto-organiser. La collaboration entre agents aide les organisations à faire évoluer leurs écosystèmes d'IA agentic avec agilité et confiance.

Intégrer la gouvernance dans un tissu agentique

Une plus grande autonomie s'accompagne d'un plus grand risque. La gouvernance doit être intégrée à l'architecture des agents dès le premier jour. Cela inclut la définition de limites politiques qui définissent ce que les agents sont autorisés à faire, l'application de modèles d'identité qui déterminent pour le compte de qui ils agissent et la mise en œuvre de l'explicabilité et de la traçabilité. Les systèmes d'observabilité doivent capturer des données télémétriques sur le comportement des agents à l'aide de services tels qu'Amazon AWS X-Ray, qui fournissent une journalisation centralisée CloudWatch et un suivi distribué sur l'ensemble des flux de travail des agents. Les agents réfléchissants peuvent continuellement auditer et évaluer les performances sur la base de ces flux de télémétrie.

La gouvernance doit également évoluer à mesure que l'écosystème des agents mûrit. À mesure que les agents gagnent en compétence et en autonomie, les mécanismes de supervision doivent

devenir plus adaptables. Les mises à jour des politiques, le blocage des capacités et les contraintes comportementales d'exécution doivent être dynamiques et applicables à grande échelle. La confiance n'est pas une caractéristique complémentaire. Il est continuellement renforcé par l'architecture, le comportement et les processus. Gestion des identités et des accès AWS (IAM) et AWS AppConfigjouent un rôle essentiel dans le renforcement des identités sécurisées, des limites d'autorisation d'exécution et des changements de comportement spécifiques à l'environnement entre les agents.

Adopter un état d'esprit opérationnel axé sur les décisions

L'automatisation traditionnelle met l'accent sur l'efficacité des processus, qui consiste à exécuter des scripts ou des flux de travail prédéfinis plus rapidement et de manière plus fiable. L'IA agentic, en revanche, introduit l'automatisation axée sur les décisions. Les agents évaluent le contexte, évaluent les options et adaptent le comportement en temps réel. Ce passage d'un état d'esprit axé sur l'exécution à un état d'esprit axé sur les décisions nécessite une nouvelle réflexion sur les indicateurs de réussite et les résultats. Au lieu de mesurer le succès uniquement en fonction de l'achèvement des tâches, le succès de l'IA agentic est mesuré par la mesure dans laquelle la décision est alignée sur l'intention, les politiques et l'évolution des conditions.

Plutôt que de mesurer uniquement l'achèvement des tâches ou le temps de cycle, les organisations doivent évaluer la qualité des décisions et la réactivité au changement. time-to-action KPIs devrait inclure des mesures telles que :

- Qualité des décisions — Dans quelle mesure l'agent a-t-il personnalisé sa réponse en fonction de l'utilisateur ou du scénario spécifique ? A-t-il pris des décisions nuancées conformes aux objectifs commerciaux et au contexte utilisateur ?
- T ime-to-action — Avec quelle rapidité et quelle intelligence l'agent a-t-il évalué une situation et y a-t-il répondu ? La latence était-elle suffisamment faible pour donner l'impression d'être adaptative et humaine ?
- Décharge cognitive — Quelle quantité d'analyses manuelles, de triage ou de prise de décisions de routine l'agent a-t-il pu gérer pour le compte d'un humain ? Cela a-t-il réduit l'effort ou l'a-t-il simplement déplacé ?

Les entreprises qui privilégient la prise de décision peuvent devenir plus résilientes, plus adaptables et capables de fonctionner à un nouveau niveau de complexité.

Évoluer en fonction du but et de l'intention

Pour réussir à faire évoluer l'IA agentic, il ne suffit pas d'expérimenter avec davantage d'outils. Il s'agit de créer une couche durable d'intelligence d'entreprise. Cela nécessite des investissements dans l'infrastructure de la plateforme, la culture opérationnelle, les cadres de gouvernance et l'alignement stratégique. Les entreprises doivent adopter une approche intentionnelle. Ils doivent traiter les agents non pas comme des expériences mais comme des éléments essentiels de leur modèle opérationnel numérique.

L'alignement sur le AWS Well-Architected Framework permet à vos systèmes de répondre aux normes de l'entreprise en matière de fiabilité, de sécurité, d'efficacité des performances et d'optimisation des coûts. Des outils tels que le SDK Strands Agents peuvent accélérer ce processus en fournissant des instructions structurées, en enregistrant les outils et en préparant les CI/CD. Cela permet aux équipes de passer de l'expérimentation à une livraison évolutive en utilisant des AWS flux de travail familiers.

L'IA agentic n'est pas un outil ; c'est un changement dans la façon dont l'intelligence est intégrée aux opérations. Organisations qui se préparent en conséquence peuvent automatiser davantage, fonctionner plus intelligemment, s'adapter plus rapidement et créer un avantage durable dans un monde de plus en plus complexe.

Conclusion pour l'opérationnalisation de l'IA agentique

L'IA agentic représente bien plus qu'un changement technologique. Il marque l'émergence d'un nouveau système d'exploitation pour l'entreprise. Organisations qui adoptent cette transformation vont au-delà des cas d'utilisation limités à l'automatisation et intègrent l'intelligence à la base de leurs opérations. Ce changement consiste à repenser la façon dont les décisions sont prises, la manière dont les systèmes s'adaptent et la manière dont les résultats sont obtenus à grande échelle.

À une époque caractérisée par une complexité croissante, une demande en temps réel et une surcharge d'informations, le modèle traditionnel d'automatisation par script a atteint ses limites. Le succès repose désormais sur la capacité à intégrer l'intelligence directement dans les flux de travail afin de créer des systèmes capables de percevoir, de raisonner, d'agir et d'évoluer. L'IA agentic peut aligner l'autonomie sur les objectifs, la prise de décision sur la gouvernance, et l'adaptabilité sur la responsabilité.

Cette transition nécessite de passer d'une approche axée sur l'exécution à une approche axée sur les décisions. Les systèmes Agentic ne se contentent pas de suivre les instructions. Ils interprètent les objectifs, évaluent les compromis et recherchent des résultats dans le cadre de contraintes définies. Dans ce contexte, le succès ne se mesure pas uniquement en fonction de l'achèvement des tâches. Il est également mesuré par la qualité, l'agilité et l'explicabilité des décisions prises en temps réel. Organisations doivent repenser les indicateurs, les incitations et la conception des systèmes afin de soutenir les agents qui agissent de manière intelligente dans des conditions d'incertitude.

L'opérationnalisation de l'IA agentique n'est pas une mise à niveau. plug-and-play Il s'agit d'une transformation architecturale et culturelle. Cela nécessite des pratiques disciplinées en matière de gestion du cycle de vie, d'application de la confiance, d'interopérabilité et d'alignement sur les modèles commerciaux. Cela nécessite également l'évolution des modèles de prestation, tels que la définition des zones d'intention, l'intégration de garde-fous en matière d'exécution et l'alignement continu du comportement des agents sur les résultats stratégiques. Les équipes doivent adopter un langage commun, une propriété partagée et une responsabilité partagée en matière de performance et de sécurité des agents.

Le niveau de préparation de l'entreprise peut déterminer qui s'épanouira dans ce nouvel environnement. Organisations doivent investir dans des cadres internes d'habilitation, de AgentOps capacités et de gouvernance capables d'évoluer et de créer de la valeur à long terme. Ceux qui réussissent peuvent créer des systèmes plus intelligents, mais ils peuvent également créer des entreprises plus adaptatives, résilientes et axées sur les connaissances.

Ce guide pose les bases. Il relie la stratégie à l'exécution et prépare les organisations à créer des plateformes évolutives d'agents intelligents. La série de contenus plus large sur l'IA agentic AWS fournit des conseils complémentaires. Pour consulter les autres guides de cette série, consultez Agentic AI sur le site Web de AWS Prescriptive Guidance. Cette série de contenus propose une feuille de route pour opérationnaliser l'autonomie avec discipline et intention.

Pour commencer, identifiez un espace décisionnel à fort impact dans lequel les agents peuvent apporter des améliorations mesurables en termes de rapidité, de précision ou de réactivité. Déployez ensuite un agent pilote ciblé doté de boucles d'instrumentation, de gouvernance et de feedback. Utilisez-le pour valider l'hypothèse de valeur, générer une dynamique interne et renforcer la confiance dans l'approche. L'élan s'amplifie grâce à l'apprentissage.

L'IA agentic n'est pas une destination ; c'est une couche de capacités qui évolue en même temps que votre entreprise. Cela représente une évolution à long terme vers le renseignement en tant qu'infrastructure. Organisations leaders dans ce domaine peuvent automatiser davantage, réagir plus rapidement, mieux s'adapter et créer des modèles opérationnels capables de gérer la complexité à l'échelle de l'entreprise.

Ressources pour opérationnaliser l'IA agentic

Services AWS

Les fonctionnalités Services AWS et fonctionnalités suivantes peuvent vous aider à créer et à mettre en œuvre des systèmes d'IA agentic dans les domaines suivants : AWS Cloud

- Amazon API Gateway peut présenter les capacités des agents comme étant évolutives et propose une tarification basée sur l'utilisation.
- AWS AppConfigpermet de gérer la configuration d'exécution et de basculer entre les fonctionnalités pour les agents entre les locataires ou les environnements.
- Amazon Bedrock est un service de base que les agents peuvent utiliser pour raisonner, générer et exécuter rapidement.
- AWS Cloud Development Kit (AWS CDK)est un service d'infrastructure sous forme de code que vous pouvez utiliser pour déployer et gérer des piles d'agents.
- AWS CloudTrailenregistre l'historique des événements afin que vous puissiez suivre l'activité des agents, les pistes d'audit et les comportements d'intégration.
- Amazon CloudWatch peut gérer les journaux, les métriques et les alarmes pour surveiller les performances des agents et le comportement de collaboration entre agents.
- AWS CodePipelinefournit une CI/CD automatisation que vous pouvez utiliser pour tester, valider et déployer le code de l'agent.
- Amazon Cognito est un service d'identité que vous pouvez utiliser pour gérer l'authentification des utilisateurs et des locataires dans les systèmes multi-agents.
- AWS Configoffre une détection de conformité et de dérive pour la politique des agents et la configuration de l'environnement.
- AWS Cost Explorerpeut suivre l'utilisation au niveau des agents et vous aider à aligner les coûts afin de maximiser votre retour sur investissement.
- Amazon DynamoDB est un service de stockage que vous pouvez utiliser pour la mémoire des agents, les journaux d'amélioration et l'état contextuel.
- Amazon Elastic File System (Amazon EFS) est un système de fichiers partagé que vous pouvez utiliser pour la collaboration des agents ou le traitement intermédiaire entre les flux de travail.
- Amazon EventBridge est un bus d'événements central que vous pouvez utiliser pour acheminer les tâches et orchestrer les communications dans la structure des agents.

- Amazon EventBridge Pipes peut rationaliser l'ingestion et le routage des événements pour connecter les agents et les services.
- Amazon GuardDuty propose des fonctionnalités de détection des menaces et de surveillance des anomalies qui peuvent garantir l'exécution sécurisée des agents.
- Gestion des identités et des accès AWS (IAM) vous aide à définir des autorisations précises pour l'exécution des agents et l'accès aux données.
- AWS Lambdaest un service informatique apatride capable d'exécuter la logique des agents et d'essaim de drones.
- AWS Marketplaceest une plateforme de distribution externe que vous pouvez utiliser pour proposer des fonctionnalités d'agent sous forme de produits commerciaux.
- AWS Organizationsest un service de gouvernance et d'application des politiques entre comptes qui peut vous aider à gérer une infrastructure d'agents multi-locataires.
- AWS Organizations les politiques de contrôle des services servent de garde-fous pour contrôler les autorisations au niveau du compte ou de l'unité organisationnelle.
- Amazon Quick est une plateforme de business intelligence (BI) générative basée sur l'IA qui vous aide à analyser les données, à créer des visualisations, à automatiser les flux de travail et à collaborer avec d'autres personnes au sein de votre organisation.
- AWS Resource Access Manager (AWS RAM) peut vous aider à partager les fonctionnalités entre les comptes et les services des agents.
- Amazon SageMaker AI est un service que vous pouvez utiliser pour la formation, le réglage précis et l'inférence des modèles au-delà des modèles de base.
- Amazon Simple Storage Service (Amazon S3) propose un stockage d'objets pour les bibliothèques rapides, les artefacts de modèles et les données générées par des agents.
- AWS Step Functionsest un moteur de flux de travail qui peut vous aider à coordonner les flux multi-agents et les pipelines de reconversion.
- AWS X-Raypropose un suivi distribué que vous pouvez utiliser pour suivre les flux de décision des agents et les dépendances des services.

Autres AWS ressources

- Les fondements de l'IA agentique sur AWS
- Modèles et flux de travail d'IA agentic sur AWS
- Frameworks, protocoles et outils d'IA agentic sur AWS

- Création d'architectures sans serveur pour l'IA agentic sur AWS
- Création d'architectures multi-locataires pour l'IA agentic sur AWS

Historique du document

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

Modification	Description	Date
Publication initiale	—	12 août 2025

AWS Glossaire des directives prescriptives

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

Nombres

7 R

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

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

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

A

A2 (1) Agent-to-Agent

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

ABAC

Voir contrôle d'accès basé sur les attributs.

services abstraits

Consultez la section Services gérés.

ACIDE

Voir atomicité, consistance, isolation, durabilité.

migration active-active

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

migration active-passive

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

Agent

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

Agent Ops

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

fonction d'agrégation

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

AI

Voir intelligence artificielle.

AIOps

Voir les opérations d'intelligence artificielle.

anonymisation

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

anti-motif

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

contrôle des applications

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

portefeuille d'applications

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

intelligence artificielle (IA)

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

opérations d'intelligence artificielle (AIOps)

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

chiffrement asymétrique

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

atomicité, cohérence, isolement, durabilité (ACID)

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

contrôle d'accès par attributs (ABAC)

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

source de données faisant autorité

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

Zone de disponibilité

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

AWS Cadre d'adoption du cloud (AWS CAF)

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

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

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

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

B

mauvais bot

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

BCP

Consultez la section Planification de la continuité des activités.

graphique de comportement

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

système de poids fort

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

classification binaire

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

filtre de Bloom

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

blue/green déploiement

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

bot

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

botnet

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

branche

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

accès par brise-vitre

Dans des circonstances exceptionnelles et par le biais d'un processus approuvé, c'est un moyen rapide pour un utilisateur d'accéder à un accès auquel Compte AWS il n'est généralement pas autorisé. Pour plus d'informations, consultez l'indicateur Mettre en œuvre des procédures permettant de briser le verre dans le AWS Well-Architected guide.

stratégie existante (brownfield)

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

cache de tampon

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

capacité métier

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

planification de la continuité des activités (BCP)

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

C

CAF

Voir le cadre d'adoption du AWS cloud.

déploiement de Canary

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

CCoE

Voir le Centre d'excellence du cloud.

CDC

Consultez la section Capture des données de modification.

capture des données de modification (CDC)

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

ingénierie du chaos

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

CI/CD

Découvrez l'intégration continue et la livraison continue.

classification

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

Développeur citoyen

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

chiffrement côté client

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

Centre d'excellence cloud (CCoE)

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

cloud computing

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

modèle d'exploitation du cloud

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

étapes d'adoption du cloud

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

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

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

CMDB

Consultez la base de données de gestion des configurations.

référentiel de code

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

cache passif

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

données gelées

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

vision par ordinateur (CV)

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

dérive de configuration

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

base de données de gestion des configurations (CMDB)

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

pack de conformité

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

intégration continue et livraison continue (CI/CD)

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

CV

Voir vision par ordinateur.

D

données au repos

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

classification des données

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

dérive des données

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

données en transit

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

maillage de données

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

minimisation des données

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

périmètre de données

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

prétraitement des données

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

provenance des données

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

sujet des données

Personne dont les données sont collectées et traitées.

entrepôt des données

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

langage de définition de base de données (DDL)

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

langage de manipulation de base de données (DML)

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

DDL

Voir langage de définition de base de données.

ensemble profond

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

deep learning

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

défense en profondeur

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

administrateur délégué

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

déploiement

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

environnement de développement

Voir environnement.

contrôle de détection

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

cartographie de la chaîne de valeur du développement (DVSM)

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

jumeau numérique

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

tableau des dimensions

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

catastrophe

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

reprise après sinistre (DR)

La stratégie et le processus que vous utilisez pour minimiser les temps d'arrêt et les pertes de données causés par un sinistre. Pour plus d'informations, consultez la section Reprise après sinistre des charges de travail sur AWS : Restauration dans le cloud dans le AWS Well-Architected Framework.

DML

Voir langage de manipulation de base de données.

conception axée sur le domaine

Approche visant à développer un système logiciel complexe en connectant ses composants à des domaines évolutifs, ou objectifs métier essentiels, que sert chaque composant. Ce concept a été introduit par Eric Evans dans son livre, Domain-Driven Design : Tackling Complexity in the Heart of Software (Boston : Addison-Wesley Professional, 2003). Pour plus d'informations sur la manière dont vous pouvez utiliser la conception axée sur le domaine avec le modèle Strangler Fig, consultez la section Modernisation incrémentielle des anciens services Web ASP.NET Microsoft (ASMX) à l'aide de conteneurs et d'Amazon API Gateway.

DR

Consultez la section Reprise après sinistre.

détection de dérive

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

DVSM

Voir la cartographie de la chaîne de valeur du développement.

E

EDA

Voir analyse exploratoire des données.

EDI

Voir échange de données informatisé.

informatique de périphérie

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

échange de données informatisé (EDI)

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

chiffrement

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

clé de chiffrement

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

endianisme

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

point de terminaison

Voir point de terminaison de service.

service de point de terminaison

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

planification des ressources d'entreprise (ERP)

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

chiffrement d'enveloppe

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

environment

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

- Environnement de développement : instance d'une application en cours d'exécution à laquelle seule l'équipe principale chargée de la maintenance de l'application peut accéder. Les environnements de développement sont utilisés pour tester les modifications avant de les promouvoir dans les environnements supérieurs. Ce type d'environnement est parfois appelé environnement de test.
- Environnements inférieurs : tous les environnements de développement d'une application, tels que ceux utilisés pour les générations et les tests initiaux.

- Environnement de production : instance d'une application en cours d'exécution à laquelle les utilisateurs finaux peuvent accéder. Dans un CI/CD pipeline, l'environnement de production est le dernier environnement de déploiement.
- Environnements supérieurs : tous les environnements accessibles aux utilisateurs autres que l'équipe de développement principale. Ils peuvent inclure un environnement de production, des environnements de préproduction et des environnements pour les tests d'acceptation par les utilisateurs.

épopée

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

ERP

Voir Planification des ressources d'entreprise.

analyse exploratoire des données (EDA)

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

F

tableau des faits

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

échouer rapidement

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

limite d'isolation des défauts

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

branche de fonctionnalités

Voir la succursale.

fonctionnalités

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

importance des fonctionnalités

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

transformation de fonctionnalité

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

invitation en quelques coups

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

FGAC

Découvrez le contrôle d'accès détaillé.

contrôle d'accès détaillé (FGAC)

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

migration instantanée (flash-cut)

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

FM

Voir le modèle de fondation.

modèle de fondation (FM)

Un vaste réseau neuronal d'apprentissage profond qui s'entraîne sur des ensembles de données massifs de données généralisées et non étiquetées. Les FM sont capables d'effectuer une grande variété de tâches générales, telles que la compréhension du langage, la génération de texte et d'images et la conversation en langage naturel. Pour plus d'informations, voir Que sont les modèles de base ?

Passerelle FM

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

G

IA générative

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

blocage géographique

Voir les restrictions géographiques.

restrictions géographiques (blocage géographique)

Sur Amazon CloudFront, option permettant d'empêcher les utilisateurs de certains pays d'accéder aux distributions de contenu. Vous pouvez utiliser une liste d'autorisation ou une liste de blocage

pour spécifier les pays approuvés et interdits. Pour plus d'informations, consultez la section Restreindre la distribution géographique de votre contenu dans la CloudFront documentation.

Flux de travail Gitflow

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

image dorée

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

stratégie inédite

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

barrière de protection

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

rambardes (AI)

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

H

HA

Découvrez la haute disponibilité.

migration de base de données hétérogène

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

haute disponibilité (HA)

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

modernisation des historiens

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

données de rétention

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

humain dans la boucle (HiTL)

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

migration de base de données homogène

Migration de votre base de données source vers une base de données cible qui partage le même moteur de base de données (par exemple, Microsoft SQL Server vers Amazon RDS for SQL Server). La migration homogène s'inscrit généralement dans le cadre d'un effort de

réhébergement ou de replateforme. Vous pouvez utiliser les utilitaires de base de données natifs pour migrer le schéma.

données chaudes

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

correctif

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

période de soins intensifs

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

I

IaC

Considérez l'infrastructure comme un code.

politique basée sur l’identité

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

application inactive

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

IIoT

Voir Internet industriel des objets.

infrastructure immuable

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

VPC entrant (d'entrée)

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

migration incrémentielle

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

Industry 4.0

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

infrastructure

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

infrastructure en tant que code (IaC)

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

internet industriel des objets (IIoT)

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

VPC d'inspection

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

Internet des objets (IoT)

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

interprétabilité

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

IoT

Voir Internet des objets.

Bibliothèque d'informations informatiques (ITIL)

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

gestion des services informatiques (ITSM)

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

ITIL

Consultez la bibliothèque d'informations informatiques.

ITSM

Voir Gestion des services informatiques.

L

contrôle d'accès basé sur des étiquettes (LBAC)

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

zone de destination

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

grand modèle de langage (LLM)

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

migration de grande envergure

Migration de 300 serveurs ou plus.

LBAC

Voir contrôle d'accès basé sur des étiquettes.

principe de moindre privilège

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

lift and shift

Voir 7 Rs.

système de poids faible

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

LLM

Voir le grand modèle de langage.

environnements inférieurs

Voir environnement.

M

machine learning (ML)

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

branche principale

Voir la succursale.

malware

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

services gérés

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

système d'exécution de la fabrication (MES)

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

MAP

Voir Migration Acceleration Program.

MCP

Voir Model Context Protocol.

Protocole de contexte du modèle (MCP)

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

serveur MCP

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

mécanisme

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

compte membre

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

MAILLES

Voir le système d'exécution de la fabrication.

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

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

microservice

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

architecture de microservices

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

Programme d'accélération des migrations (MAP)

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

migration à grande échelle

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

usine de migration

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

métadonnées de migration

Informations relatives à l'application et au serveur nécessaires pour finaliser la migration. Chaque modèle de migration nécessite un ensemble de métadonnées de migration différent. Les

exemples de métadonnées de migration incluent le sous-réseau cible, le groupe de sécurité et le AWS compte.

modèle de migration

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

Évaluation du portefeuille de migration (MPA)

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

Évaluation de la préparation à la migration (MRA)

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

stratégie de migration

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

ML

Voir apprentissage automatique.

modernisation

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

évaluation de la préparation à la modernisation

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

applications monolithiques (monolithes)

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

MPA

Voir Évaluation du portefeuille de migration.

MQTT

Voir Message Queuing Telemetry Transport.

classification multi-classes

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

infrastructure mutable

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

O

OAC

Voir Contrôle d'accès à l'origine.

OAI

Voir l'identité d'accès à l'origine.

OCM

Voir gestion du changement organisationnel.

migration hors ligne

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

OI

Consultez la section Intégration des opérations.

OLA

Voir l'accord au niveau opérationnel.

migration en ligne

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

OPC-UA

Voir Open Process Communications - Architecture unifiée.

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

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

accord au niveau opérationnel (OLA)

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

examen de l'état de préparation opérationnelle (ORR)

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

technologie opérationnelle (OT)

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

intégration des opérations (OI)

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

journal de suivi d'organisation

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

gestion du changement organisationnel (OCM)

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

contrôle d'accès d'origine (OAC)

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

identité d'accès d'origine (OAI)

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

ORR

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

DE

Voir technologie opérationnelle.

VPC sortant (de sortie)

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

P

limite des autorisations

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

informations personnelles identifiables (PII)

Informations qui, lorsqu'elles sont consultées directement ou associées à d'autres données connexes, peuvent être utilisées pour déduire raisonnablement l'identité d'une personne. Les

exemples d'informations personnelles incluent les noms, les adresses et les informations de contact.

PII

Voir les informations personnelles identifiables.

manuel stratégique

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

PLC

Voir contrôleur logique programmable.

PLM

Consultez la section Gestion du cycle de vie des produits.

policy

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

persistance polyglotte

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

évaluation du portefeuille

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

predicate

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

prédicat pushdown

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

contrôle préventif

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

principal

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

confidentialité dès la conception

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

zones hébergées privées

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

contrôle proactif

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

gestion du cycle de vie des produits (PLM)

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

environnement de production

Voir environnement.

contrôleur logique programmable (PLC)

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

chaînage rapide

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

pseudonymisation

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

publish/subscribe (pub/sub)

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

Q

plan de requête

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

régression du plan de requêtes

Le cas où un optimiseur de service de base de données choisit un plan moins optimal qu'avant une modification donnée de l'environnement de base de données. Cela peut être dû à des

changements en termes de statistiques, de contraintes, de paramètres d'environnement, de liaisons de paramètres de requêtes et de mises à jour du moteur de base de données.

R

Matrice RACI

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

RAG

Voir Retrieval Augmented Generation.

rançongiciel

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

Matrice RASCI

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

RCAC

Voir contrôle d'accès aux lignes et aux colonnes.

réplica en lecture

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

réarchitecte

Voir 7 Rs.

objectif de point de récupération (RPO)

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

objectif de temps de récupération (RTO)

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

refactoriser

Voir 7 Rs.

Région

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

régression

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

réhéberger

Voir 7 Rs.

version

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

déplacer

Voir 7 Rs.

replateforme

Voir 7 Rs.

rachat

Voir 7 Rs.

résilience

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

politique basée sur les ressources

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

matrice responsable, redevable, consulté et informé (RACI)

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

contrôle réactif

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

retain

Voir 7 Rs.

se retirer

Voir 7 Rs.

Génération augmentée de récupération (RAG)

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

rotation

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

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

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

RPO

Voir l'objectif du point de récupération.

RTO

Voir l'objectif en matière de temps de rétablissement.

runbook

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

S

SAML 2.0

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

SCADA

Voir Contrôle de supervision et acquisition de données.

SCP

Voir la politique de contrôle des services.

secret

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

sécurité dès la conception

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

contrôle de sécurité

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

renforcement de la sécurité

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

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

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

automatisation des réponses de sécurité

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

chiffrement côté serveur

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

Politique de contrôle des services (SCP)

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

point de terminaison du service

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

contrat de niveau de service (SLA)

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

indicateur de niveau de service (SLI)

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

objectif de niveau de service (SLO)

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

modèle de responsabilité partagée

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

IA de l'ombre

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

SIEM

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

point de défaillance unique (SPOF)

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

SLA

Voir le contrat de niveau de service.

SLI

Voir l'indicateur de niveau de service.

SLO

Voir l'objectif de niveau de service.

modèle split-and-seed

Modèle permettant de mettre à l'échelle et d'accélérer les projets de modernisation. Au fur et à mesure que les nouvelles fonctionnalités et les nouvelles versions de produits sont définies, l'équipe principale se divise pour créer des équipes de produit. Cela permet de mettre à l'échelle

les capacités et les services de votre organisation, d'améliorer la productivité des développeurs et de favoriser une innovation rapide. Pour plus d'informations, consultez la section Approche progressive de la modernisation des applications dans le. AWS Cloud

SPOF

Voir point de défaillance unique.

schéma en étoile

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

modèle de figuier étrangleur

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

sous-réseau

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

contrôle de supervision et acquisition de données (SCADA)

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

chiffrement symétrique

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

tests synthétiques

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

invite du système

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

T

tags

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

variable cible

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

liste de tâches

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

environnement de test

Voir environnement.

entraînement

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

outil

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

passerelle de transit

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

flux de travail basé sur jonction

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

accès sécurisé

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

réglage

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

équipe de deux pizzas

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

U

incertitude

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

tâches indifférenciées

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

environnements supérieurs

Voir environnement.

V

mise à vide

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

contrôle de version

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

Appairage de VPC

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

vulnérabilités

Défaut logiciel ou matériel qui compromet la sécurité du système.

W

cache actif

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

données chaudes

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

fonction de fenêtre

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

charge de travail

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

flux de travail

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

VER

Voir écrire une fois, lire plusieurs.

WQF

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

écrire une fois, lire plusieurs (WORM)

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

Z

exploit Zero-Day

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

vulnérabilité de type « jour zéro »

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

invite Zero-Shot

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

application zombie

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

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