Frameworks, plateformes, protocoles et outils d'IA agentic sur AWS

AWS Conseils prescriptifs

AWS Conseils prescriptifs: Frameworks, plateformes, protocoles et outils d'IA agentic sur AWS

Table of Contents

Frameworks, plateformes, protocoles et outils d'IA agentic sur AWS

Aaron Sempf, Ansley Verzosa et Joshua Samuel, Amazon Web Services ()AWS

Janvier 2026 (historique du document)

L'IA agentique est un puissant paradigme à l'intersection de l'IA, des systèmes distribués et du génie logiciel. Il s'agit d'une classe de systèmes intelligents composés d'agents logiciels autonomes et asynchrones qui utilisent des modèles d'IA et s'intègrent à des outils et à des ressources. Les agents font preuve d'agentivité, peuvent percevoir le contexte, raisonner plutôt que les objectifs, prendre des décisions et prendre des mesures ciblées au nom des utilisateurs ou des systèmes. Ces agents opèrent de manière indépendante, souvent en collaboration, dans des environnements distribués et sont conçus pour poursuivre des objectifs délégués grâce à l'intelligence, à la mémoire et à l'intention intégrées.

En effet AWS, les entreprises peuvent tirer parti de l'IA agentic pour automatiser des flux de travail complexes, améliorer les processus de prise de décision et créer des systèmes plus réactifs.
Ce guide fournit des informations sur les composants clés nécessaires pour créer des solutions d'intelligence artificielle agentiques efficaces :

- Les frameworks dressent le profil des frameworks d'IA agentic actuels, y compris des examens de leurs avantages et de leurs cas d'utilisation. Découvrez comment ces cadres réduisent le transport indifférencié de charges lourdes selon les modèles, les protocoles et les outils. Comprenez les principaux critères de sélection afin de choisir le cadre adapté à vos besoins.
- Platforms fournit une vue d'ensemble des plateformes d'intelligence artificielle agentique (agent géré, orchestration open source et hybride) et des considérations relatives à la sélection ou à la conception.
- Protocoles explore les protocoles de communication standardisés essentiels pour les interactions entre agents. Agent-to-agentdes protocoles émergent, tels que le Model Context Protocol (MCP) open source et Agent2Agent (A2A), ainsi que d'autres implémentations propriétaires. Découvrez comment les protocoles communs permettent aux différents protocoles d'interagir de manière fluide.
- Tools fournit des informations sur les outils basés sur des protocoles (tels que le MCP), les outils natifs du framework et les méta-outils. Organisations peuvent créer une boîte à outils qui s'intègre

aux principaux systèmes de leurs flux de travail, permettant à la fois aux utilisateurs finaux et aux flux de travail agentiques basés sur le serveur.

Public visé

Ce guide s'adresse aux architectes, aux développeurs et aux leaders technologiques qui cherchent à exploiter la puissance des agents logiciels pilotés par l'IA au sein d'applications cloud natives modernes.

Objectifs

Ce guide vous aide à accomplir les tâches suivantes :

- Comparez différents frameworks d'IA agentic pour sélectionner celui qui convient le mieux à votre cas d'utilisation.
- Découvrez les plateformes d'IA agentic qui fournissent des capacités permettant de transformer des agents individuels en systèmes coordonnés et adaptatifs.
- Découvrez les avantages des protocoles ouverts pour créer des architectures d'IA agentiques durables.
- Créez une stratégie d'intégration d'outils appropriée lors de la création de systèmes d'agents.

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.

Cadres

Les fondements de l'IA agentic AWS examinent les principaux modèles et flux de travail qui permettent un comportement autonome et axé sur les objectifs. Le choix du cadre est au cœur de la mise en œuvre de ces modèles. Un framework est la base logicielle du code préécrit qui fournit un environnement structuré et des fonctionnalités communes pour la création et la gestion, les outils et les capacités d'orchestration nécessaires pour créer des agents d'IA autonomes prêts pour la production.

Les frameworks d'IA agentique efficaces fournissent plusieurs fonctionnalités essentielles qui transforment les interactions brutes des grands modèles de langage (LLM) en systèmes coordonnés et intelligents capables de raisonner, de collaborer et d'agir :

- L'orchestration des agents coordonne le flux d'informations et la prise de décision entre un ou plusieurs agents afin d'atteindre des objectifs complexes sans intervention humaine.
- L'intégration d'outils permet aux agents d'interagir avec des systèmes externes et des sources de données afin d'étendre leurs capacités au-delà du traitement du langage. APIs Pour plus d'informations, consultez la section Présentation des outils dans la Strands Agents documentation.
- La gestion de la mémoire fournit un état persistant ou basé sur les sessions afin de maintenir le contexte dans toutes les interactions, ce qui est essentiel pour les tâches de longue durée ou adaptatives. Des frameworks plus avancés intègrent une mémoire à long terme pour stocker les résumés et les préférences des utilisateurs, permettant ainsi des expériences agentiques personnalisées et adaptées au contexte. Pour plus d'informations, consultez la section Comment penser aux frameworks d'agents sur le LangChain blog.
- La définition du flux de travail prend en charge des modèles structurés tels que les chaînes, le routage, la parallélisation et les boucles de réflexion qui permettent un raisonnement autonome sophistiqué.
- Le déploiement et le suivi facilitent le passage du développement à la production grâce à l'observabilité pour les systèmes autonomes. Pour plus d'informations, consultez l'annonce de disponibilité AgentCore générale d'Amazon Bedrock.

Ces fonctionnalités sont mises en œuvre selon différentes approches et approches dans l'ensemble du paysage des frameworks, chacune offrant des avantages distincts pour différents cas d'utilisation d'agents autonomes et contextes organisationnels.

Cette section décrit et compare les principaux frameworks pour la création de solutions d'IA agentiques, en mettant l'accent sur leurs points forts, leurs limites et leurs cas d'utilisation idéaux pour un fonctionnement autonome :

- Agents à mèches
- LangChain et LangGraph
- Équipage AI
- AutoGen
- ???
- Comparaison des frameworks d'IA agentiques

> **Note**
>
> Cette section couvre les cadres qui soutiennent spécifiquement l'agence de l'IA et ne couvre pas les interfaces frontales ou l'IA générative sans agence.

Strands Agents

Strands Agentsest un SDK open source initialement publié par AWS, comme décrit dans le blog AWS Open Source. Strands Agentsest conçu pour créer des agents d'intelligence artificielle autonomes selon une approche axée sur le modèle. Il fournit un cadre flexible et extensible conçu pour fonctionner parfaitement Services AWS tout en restant ouvert à l'intégration avec des composants tiers. Strands Agents est idéal pour créer des solutions totalement autonomes.

Principales fonctionnalités de Strands Agents

Strands Agentsinclut les principales fonctionnalités suivantes :

- Conception axée sur le modèle : construite autour du concept selon lequel le modèle de base est au cœur de l'intelligence des agents, permettant un raisonnement autonome sophistiqué. Pour plus d'informations, consultez Agent Loop dans la Strands Agents documentation.
- Modèles de collaboration multi-agents : modèles de coordination intégrés tels que les modèles Swarm, Graph et Workflow qui permettent une collaboration et une gouvernance évolutives sur les réseaux d'agents distribués. Pour plus d'informations, consultez la section Modèles multi-agents dans la documentation Strands Agents.

- Intégration MCP : support natif du protocole MCP (Model Context Protocol), permettant une fourniture de contexte standardisée LLMs pour un fonctionnement autonome cohérent.
- Service AWS intégration — Connexion fluide à Amazon Bedrock,, AWS Lambda AWS Step Functions, et autres Services AWS pour des flux de travail autonomes complets. Pour plus d'informations, consultez le résuméAWS hebdomadaire (AWS blog).
- Sélection du modèle de base : prend en charge différents modèles de base, notamment Anthropic Claude, Amazon Nova (Premier, Pro, Lite et Micro) sur Amazon Bedrock, et d'autres pour optimiser les différentes capacités de raisonnement autonome. Pour plus d'informations, consultez Amazon Bedrock dans la Strands Agents documentation.
- Intégration de l'API LLM — Intégration flexible avec différentes interfaces de service LLM, notamment Amazon Bedrock, OpenAI, etc. pour le déploiement en production. Pour plus d'informations, consultez Amazon Bedrock Basic Usage dans la Strands Agents documentation.
- Capacités multimodales — Support de plusieurs modalités, notamment le traitement du texte, de la parole et de l'image pour des interactions complètes avec les agents autonomes. Pour plus d'informations, consultez Amazon Bedrock Multimodal Support dans la Strands Agents documentation.
- Écosystème d'outils : ensemble complet d'outils d' Service AWS interaction, avec extensibilité pour les outils personnalisés qui étendent les capacités autonomes. Pour plus d'informations, consultez la section Présentation des outils dans la Strands Agents documentation.

Quand utiliser Strands Agents

Strands Agentsest particulièrement bien adapté aux scénarios d'agents autonomes, notamment :

- Organisations qui s'appuient sur une AWS infrastructure et qui souhaitent une intégration native Services AWS pour des flux de travail autonomes
- Équipes nécessitant des fonctionnalités de sécurité, d'évolutivité et de conformité de niveau professionnel pour les systèmes autonomes de production
- Projets nécessitant une flexibilité dans la sélection de modèles entre différents fournisseurs pour des tâches autonomes spécialisées
- Cas d'utilisation nécessitant une intégration étroite avec les AWS flux de travail et les ressources existants pour des processus autonomes de bout en bout

Approche de mise en œuvre pour Strands Agents

Strands Agentspropose une approche de mise en œuvre simple pour les parties prenantes de l'entreprise, comme indiqué dans son guide de démarrage rapide. Le cadre permet aux organisations de :

- Sélectionnez des modèles de base tels qu'Amazon Nova (Premier, Pro, Lite ou Micro) sur Amazon Bedrock en fonction des besoins commerciaux spécifiques.
- Définissez des outils personnalisés qui se connectent aux systèmes et aux sources de données de l'entreprise.
- Traitez plusieurs modalités, notamment le texte, les images et le discours.
- Déployez des agents capables de répondre de manière autonome aux requêtes commerciales et d'effectuer des tâches.

Cette approche de mise en œuvre permet aux équipes commerciales de développer et de déployer rapidement des agents autonomes sans expertise technique approfondie dans le développement de modèles d'IA.

Exemple concret de Strands Agents

AWS Transform pour .NET utilise Strands Agents ses capacités de modernisation des applications, comme décrit dans AWS Transform for .NET, le premier service d'intelligence artificielle agentique permettant de moderniser les applications .NET à grande échelle (AWS blog). Ce service de production emploie plusieurs agents autonomes spécialisés. Les agents travaillent ensemble pour analyser les applications .NET existantes, planifier des stratégies de modernisation et exécuter des transformations de code vers des architectures cloud natives sans intervention humaine. AWS Transform for .NET démontre l'état de préparation à la production Strands Agents des systèmes autonomes d'entreprise.

LangChain et LangGraph

LangChainest l'un des frameworks les plus établis de l'écosystème de l'IA agentique. LangGraphétend ses fonctionnalités pour prendre en charge les flux de travail complexes et dynamiques des agents, comme décrit dans le LangChainblog. Ensemble, ils fournissent une solution complète pour créer des agents IA autonomes sophistiqués dotés de riches capacités d'orchestration pour un fonctionnement indépendant.

Principales caractéristiques de LangChain et LangGraph

LangChainet LangGraph incluent les fonctionnalités clés suivantes :

- Écosystème de composants — Vaste bibliothèque de composants prédéfinis pour diverses fonctionnalités d'agents autonomes, permettant le développement rapide d'agents spécialisés. Pour plus d'informations, consultez Quickstart dans la LangChain documentation.
- Sélection du modèle de base — Support pour divers modèles de fondation, notamment les modèles Anthropic Claude, Amazon Nova (Premier, Pro, Lite et Micro) sur Amazon Bedrock, et d'autres pour différentes capacités de raisonnement. Pour plus d'informations, consultez la section Entrées et sorties dans la LangChain documentation.
- Intégration de l'API LLM : interfaces standardisées pour plusieurs fournisseurs de services de grands modèles linguistiques (LLM), notamment Amazon BedrockOpenAI, pour un déploiement flexible. Pour plus d'informations, consultez la section LLMsdans la documentation LangChain.
- Traitement multimodal : prise en charge intégrée du traitement du texte, de l'image et du son pour permettre des interactions multimodales riches entre agents autonomes. Pour plus d'informations, consultez la section Multimodalité dans la LangChain documentation.
- Flux de travail basés sur des graphes : LangGraph permettent de définir les comportements complexes des agents autonomes en tant que machines à états, prenant en charge une logique décisionnelle sophistiquée. Pour plus d'informations, consultez l'annonce de LangGraphPlatform GA.
- Abstractions de mémoire — Plusieurs options pour la gestion de la mémoire à court et à long terme, ce qui est essentiel pour les agents autonomes qui maintiennent le contexte au fil du temps. Pour plus d'informations, consultez Comment ajouter de la mémoire aux chatbots dans la LangChain documentation.
- Intégration d'outils — Écosystème riche d'intégrations d'outils à travers divers services et extension APIs des capacités des agents autonomes. Pour plus d'informations, consultez la section Outils de la LangChain documentation.
- LangGraph plateforme — Solution gérée de déploiement et de surveillance pour les environnements de production, prenant en charge les agents autonomes de longue durée. Pour plus d'informations, consultez l'annonce de LangGraphPlatform GA.

Quand utiliser LangChain et LangGraph

LangChainet LangGraph sont particulièrement bien adaptés aux scénarios d'agents autonomes, notamment :

- Workflows de raisonnement complexes en plusieurs étapes qui nécessitent une orchestration sophistiquée pour une prise de décision autonome
- Projets nécessitant l'accès à un vaste écosystème de composants prédéfinis et d'intégrations pour diverses capacités autonomes
- Équipes disposant d'une infrastructure et d'une expertise en Python apprentissage automatique (ML) existantes et souhaitant créer des systèmes autonomes
- Cas d'utilisation nécessitant une gestion d'état complexe lors de sessions d'agents autonomes de longue durée

Approche de mise en œuvre pour LangChain et LangGraph

LangChainet LangGraph fournir une approche de mise en œuvre structurée pour les parties prenantes de l'entreprise, comme indiqué dans la LangGraphdocumentation. Le cadre permet aux organisations de :

- Définissez des graphiques de flux de travail sophistiqués qui représentent les processus métier.
- Créez des modèles de raisonnement en plusieurs étapes avec des points de décision et une logique conditionnelle.
- Intégrez des capacités de traitement multimodales pour gérer divers types de données.
- Mettez en œuvre le contrôle qualité grâce à des mécanismes de révision et de validation intégrés.

Cette approche basée sur des graphiques permet aux équipes commerciales de modéliser des processus décisionnels complexes sous forme de flux de travail autonomes. Les équipes ont une visibilité claire sur chaque étape du processus de raisonnement et sont en mesure d'auditer les parcours décisionnels.

Exemple concret de et LangChain LangGraph

Vodafonea mis en place des agents autonomes utilisant LangChain (etLangGraph) pour améliorer ses flux de travail d'ingénierie des données et d'exploitation, comme indiqué dans son étude de cas sur l'LangChainentreprise. Ils ont créé des assistants IA internes qui surveillent de manière

autonome les indicateurs de performance, extraient des informations des systèmes de documentation et présentent des informations exploitables, le tout par le biais d'interactions en langage naturel.

La Vodafone mise en œuvre utilise des chargeurs de documents LangChain modulaires, l'intégration vectorielle et la prise en charge de plusieurs LLMs (OpenAI, LLaMA 3 etGemini) pour prototyper et comparer rapidement ces pipelines. Ils ont ensuite structuré LangGraph l'orchestration multi-agents en déployant des sous-agents modulaires. Ces agents exécutent des tâches de collecte, de traitement, de synthèse et de raisonnement. LangGraphont intégré ces agents APIs dans leurs systèmes cloud.

CrewAI

CrewAIest un framework open source spécifiquement axé sur l'orchestration multi-agents autonome, disponible sur. GitHub Il propose une approche structurée pour créer des équipes d'agents autonomes spécialisés qui collaborent pour résoudre des tâches complexes sans intervention humaine. CrewAImet l'accent sur la coordination basée sur les rôles et la délégation des tâches.

Principales fonctionnalités de CrewAI

CrewAIfournit les fonctionnalités clés suivantes :

- Conception d'agents basée sur les rôles — Les agents autonomes sont définis avec des rôles, des objectifs et des histoires spécifiques afin de permettre une expertise spécialisée. Pour plus d'informations, consultez la section Création d'agents efficaces dans la CrewAI documentation.
- Délégation de tâches : mécanismes intégrés permettant d'attribuer des tâches de manière autonome aux agents appropriés en fonction de leurs capacités. Pour plus d'informations, consultez la section Tâches de la CrewAI documentation.
- Collaboration entre agents : cadre pour la communication autonome entre agents et le partage des connaissances sans médiation humaine. Pour plus d'informations, consultez la section Collaboration dans la CrewAI documentation.
- Gestion des processus : flux de travail structurés pour l'exécution séquentielle et parallèle de tâches autonomes. Pour plus d'informations, consultez la section Processus dans la CrewAI documentation.
- Sélection du modèle de base — Support de différents modèles de base, notamment les modèles Anthropic Claude, Amazon Nova (Premier, Pro, Lite et Micro) sur Amazon Bedrock, et d'autres pour optimiser les différentes tâches de raisonnement autonomes. Pour plus d'informations, consultez la section LLMsdans la documentation CrewAI.

- Intégration de l'API LLM — Intégration flexible avec plusieurs interfaces de service LLM, notamment Amazon BedrockOpenAI, et les déploiements de modèles locaux. Pour plus d'informations, consultez les exemples de configuration des fournisseurs dans la CrewAI documentation.
- Support multimodal — Capacités émergentes de gestion du texte, des images et d'autres modalités pour des interactions complètes avec les agents autonomes. Pour plus d'informations, consultez la section Utilisation d'agents multimodaux dans la CrewAI documentation.

Quand utiliser CrewAI

CrewAIest particulièrement bien adapté aux scénarios d'agents autonomes, notamment :

- Problèmes complexes bénéficiant d'une expertise spécialisée basée sur les rôles travaillant de manière autonome
- Projets nécessitant une collaboration explicite entre plusieurs agents autonomes
- Cas d'utilisation où la décomposition des problèmes en équipe améliore la résolution autonome des problèmes
- Scénarios nécessitant une séparation claire des préoccupations entre les différents rôles d'agent autonome

Approche de mise en œuvre pour CrewAI

CrewAIfournit une implémentation basée sur les rôles de l'approche des équipes d'agents IA pour les parties prenantes de l'entreprise, comme indiqué dans la section Getting Started de la CrewAI documentation. Le cadre permet aux organisations de :

- Définissez des agents autonomes spécialisés dotés de rôles, d'objectifs et de domaines d'expertise spécifiques.
- Attribuez des tâches aux agents en fonction de leurs capacités spécialisées.
- Établissez des dépendances claires entre les tâches pour créer des flux de travail structurés.
- Orchestrez la collaboration entre plusieurs agents pour résoudre des problèmes complexes.

Cette approche basée sur les rôles reflète les structures humaines des équipes, ce qui la rend intuitive à comprendre et à mettre en œuvre pour les chefs d'entreprise. Organisations peuvent

créer des équipes autonomes dotées de domaines d'expertise spécialisés qui collaborent pour atteindre leurs objectifs commerciaux, de la même manière que les équipes humaines fonctionnent. Cependant, l'équipe autonome peut travailler en continu sans intervention humaine.

Exemple concret de CrewAI

AWS a mis en œuvre des systèmes multi-agents autonomes utilisant CrewAI intégré à Amazon Bedrock, comme indiqué dans l'CrewAIétude de cas publiée. AWS et CrewAI a développé un cadre sécurisé et indépendant du fournisseur. L'architecture CrewAI open source « flows-and-crews » s'intègre parfaitement aux modèles de base, aux systèmes de mémoire et aux dispositifs de conformité d'Amazon Bedrock.

Les principaux éléments de la mise en œuvre sont les suivants :

- Des plans et des sources ouvertes, AWS ainsi que des modèles de référence CrewAI publiés qui associent les CrewAI agents aux modèles et aux outils d'observabilité d'Amazon Bedrock. Ils ont également publié des systèmes exemplaires tels qu'une équipe d'audit de AWS sécurité multi-agents, des flux de modernisation du code et l'automatisation du back-office des biens de consommation (CPG).
- Intégration de la pile d'observabilité : la solution intègre la surveillance avec Amazon CloudWatch et permet la traçabilité et le débogageLangFuse, de la validation du concept à la production. AgentOps
- Retour sur investissement (ROI) démontré — Les premiers projets pilotes présentent des améliorations majeures : exécution 70 % plus rapide pour un projet de modernisation du code de grande envergure et réduction d'environ 90 % du temps de traitement pour un flux de backoffice CPG.

AutoGen

AutoGenest un framework open source initialement publié parMicrosoft. AutoGense concentre sur la mise en place d'agents IA autonomes conversationnels et collaboratifs. Il fournit une architecture flexible pour créer des systèmes multi-agents en mettant l'accent sur les interactions asynchrones et pilotées par des événements entre les agents pour des flux de travail autonomes complexes.

Principales fonctionnalités de AutoGen

AutoGenfournit les fonctionnalités clés suivantes :

- Agents conversationnels : conçus autour de conversations en langage naturel entre agents autonomes, permettant un raisonnement sophistiqué par le biais du dialogue. Pour plus d'informations, consultez la section Multi-agent Conversation Framework dans la AutoGen documentation.
- Architecture asynchrone : conception axée sur les événements pour des interactions non bloquantes avec des agents autonomes, prenant en charge des flux de travail parallèles complexes. Pour plus d'informations, consultez la section Résolution de plusieurs tâches dans une séquence de discussions asynchrones dans la AutoGen documentation.
- H uman-in-the-loop — Soutien solide à la participation humaine facultative à des flux de travail d'agents par ailleurs autonomes en cas de besoin. Pour plus d'informations, consultez la section Autorisation du feedback humain dans les agents dans la AutoGen documentation.
- Génération et exécution de code — Fonctionnalités spécialisées pour les agents autonomes axés sur le code qui peuvent écrire et exécuter du code. Pour plus d'informations, consultez la section Exécution du code dans la AutoGen documentation.
- Comportements personnalisables — Configuration flexible des agents autonomes et contrôle des conversations pour divers cas d'utilisation. Pour plus d'informations, consultez agentchat.conversable_agent dans la documentation. AutoGen
- Sélection du modèle de base — Support pour différents modèles de base, notamment les modèles Anthropic Claude, Amazon Nova (Premier, Pro, Lite et Micro) sur Amazon Bedrock, et d'autres pour différentes capacités de raisonnement autonome. Pour plus d'informations, consultez la section Configuration LLM dans la AutoGen documentation.
- Intégration de l'API LLM — Configuration standardisée pour plusieurs interfaces de service LLM, notamment Amazon Bedrock etOpenAI. Azure OpenAI Pour plus d'informations, consultez oai.openai_utils dans le manuel de référence des API. AutoGen
- Traitement multimodal — Support du traitement du texte et de l'image pour permettre des interactions multimodales riches entre agents autonomes. Pour plus d'informations, consultez la section Engagement avec les modèles multimodaux : GPT-4V dans AutoGen la documentation. AutoGen

Quand utiliser AutoGen

AutoGenest particulièrement bien adapté aux scénarios d'agents autonomes, notamment :

- Applications qui nécessitent des flux conversationnels naturels entre agents autonomes pour un raisonnement complexe

- Projets nécessitant à la fois un fonctionnement totalement autonome et des capacités de supervision humaine optionnelles
- Cas d'utilisation impliquant la génération, l'exécution et le débogage de code autonomes sans intervention humaine
- Scénarios nécessitant des modèles de communication flexibles et asynchrones avec des agents autonomes

Approche de mise en œuvre pour AutoGen

AutoGenpropose une approche de mise en œuvre conversationnelle pour les parties prenantes de l'entreprise, comme indiqué dans la section Getting Started de la AutoGen documentation. Le cadre permet aux organisations de :

- Créez des agents autonomes qui communiquent par le biais de conversations en langage naturel.
- Implémentez des interactions asynchrones pilotées par des événements entre plusieurs agents.
- Combinez un fonctionnement entièrement autonome avec une supervision humaine optionnelle en cas de besoin.
- Développez des agents spécialisés pour différentes fonctions commerciales qui collaborent par le biais du dialogue.

Cette approche conversationnelle rend le raisonnement du système autonome transparent et accessible aux utilisateurs professionnels. Les décideurs peuvent observer le dialogue entre les agents pour comprendre comment les conclusions sont tirées et éventuellement participer à la conversation lorsque le jugement humain est requis.

Exemple concret de AutoGen

Magentic-Oneest un système multi-agents généraliste open source conçu pour résoudre de manière autonome des tâches complexes en plusieurs étapes dans divers environnements, comme décrit dans le blog AI Frontiers. Microsoft À la base se trouve l'agent Orchestrator, qui décompose les objectifs de haut niveau et suit les progrès à l'aide de registres structurés. Cet agent délègue des sous-tâches à des agents spécialisés (tels queWebSurfer, FileSurferCoder, etComputerTerminal) et s'adapte dynamiquement en replanifiant si nécessaire.

Le système repose sur le AutoGen framework et est indépendant du modèle, avec GPT-4o par défaut. Il atteint des performances de pointe sur des critères tels GAIA que, et, le tout sans réglage

spécifique à une tâche. AssistantBench WebArena De plus, il prend en charge l'extensibilité modulaire et une évaluation rigoureuse par le biais de AutoGenBench suggestions.

LlamaIndex

LlamaIndexest un framework de données conçu spécifiquement pour connecter de grands modèles linguistiques (LLMs) à des sources de données externes afin de permettre des applications sophistiquées de génération augmentée de récupération (RAG) et d'intelligence artificielle agentique. Le framework fournit des abstractions et des flux de développement accélérés pour les systèmes agentiques, des modèles d'orchestration personnalisés et des intégrations de systèmes qui réduisent time-to-production le nombre de solutions d'IA axées sur les connaissances.

Principales fonctionnalités de LlamaIndex

LlamaIndexfournit un ensemble complet de fonctionnalités qui le rendent particulièrement adapté aux applications d'intelligence artificielle agentiques d'entreprise :

- Architecture centrée sur les données : excelle dans l'ingestion, l'indexation et la récupération d'informations provenant de plus de 100 formats de données PDFs, notamment des documents Microsoft Word, des feuilles de calcul, etc. Le framework transforme les données d'entreprise en bases de connaissances consultables optimisées pour les agents d'intelligence artificielle. Pour plus d'informations, consultez la documentation LlamaIndex.
- Déploiement prêt pour la production : LlamaIndex propose à la fois des frameworks open source et des services gérésLlamaCloud, fournissant des fonctionnalités de niveau entreprise, notamment des contrôles de sécurité, une évolutivité, des intégrations d'observabilité et une flexibilité de déploiement. Pour plus d'informations, consultez la documentation du LlamaIndex framework.
- Traitement avancé des documents : LlamaCloud fournit des fonctionnalités d'analyse, d'extraction, d'indexation et de récupération de documents qui gèrent les mises en page complexes, les tableaux imbriqués, le contenu multimodal et même les notes manuscrites. Cette analyse sophistiquée permet aux agents de travailler efficacement avec des documents d'entreprise réels contenant des graphiques, des diagrammes et des mises en forme complexes. Pour plus d'informations, consultez la documentation LlamaCloud.
- Orchestration des flux de travail : LlamaAgents fournit un moteur d'orchestration asynchrone piloté par les événements pour créer des systèmes agentiques en plusieurs étapes. Les flux de travail prennent en charge des modèles complexes tels que les boucles, l'exécution parallèle, le branchement conditionnel et la reprise dynamique, ce qui les rend idéaux pour les interactions

sophistiquées avec les agents. Pour plus d'informations, consultez la documentation sur les LlamaIndex flux de travail.

- Capacités de récupération agentique : modes de récupération avancés, notamment la recherche hybride, la recherche sémantique et le routage automatique, qui déterminent intelligemment la meilleure stratégie de récupération pour chaque requête. Le framework prend en charge la récupération composite dans plusieurs bases de connaissances avec un reclassement pour une précision accrue. Pour plus d'informations, consultez la documentation LlamaIndex RAG.
- Observabilité et évaluation : LlamaIndex s'intègre à une variété d'outils d'observabilité et d'évaluation. Cette fonctionnalité d'intégration vous permet de suivre et de déboguer vos applications, d'évaluer leurs performances et de surveiller les coûts. Pour plus d'informations, consultez la LlamaIndex documentation relative au suivi, au débogage et à l'évaluation.

Quand utiliser LlamaIndex

LlamaIndexest particulièrement bien adapté aux scénarios d'IA agentique qui mettent l'accent sur les flux de travail gourmands en données et la gestion des connaissances :

- Applications gourmandes en documents qui nécessitent que les agents traitent, analysent et extraient des informations à partir de grands volumes de documents d'entreprise tels que des contrats, des rapports, des manuels et des documents réglementaires
- Scénarios du prototypage rapide à la production dans lesquels les entreprises souhaitent créer et déployer rapidement des agents centrés sur les documents sans surcharger la gestion de l'infrastructure
- Des architectures Rag-first qui privilégient la précision de l'extraction et la pertinence du contexte, en particulier lorsque vous travaillez avec des documents multimodaux complexes contenant des tableaux, des images et des données structurées
- Des flux de travail documentaires multi-agents qui nécessitent des agents spécialisés pour différents aspects du traitement des documents, tels que l'analyse, la synthèse et le contrôle de conformité

Approche de mise en œuvre pour LlamaIndex

LlamaIndex fournit à la fois des éléments de base de base et des abstractions de haut niveau qui s'adaptent à différentes approches de mise en œuvre :

- Développement rapide d'applications RAG fonctionnelles en quelques lignes de code grâce à une utilisation de LlamaIndex haut niveau APIs. Cette approche rend LlamaIndex accessible aux équipes commerciales et aux développeurs novices en matière d'IA agentic.
- Intégration d'entreprise via LlamaHub les systèmes d'entreprise les plus courants SharePoint, notamment Amazon Simple Storage Service (Amazon S3), les bases de données et. APIs Cette approche permet une intégration parfaite avec l'infrastructure de données existante.
- Des options de déploiement flexibles entre des déploiements open source auto-hébergés pour un contrôle maximal ou des services LlamaCloud gérés pour réduire les frais opérationnels et les fonctionnalités d'entreprise.
- Les applications peuvent commencer par de simples moteurs de requêtes et ajouter progressivement des fonctionnalités agentiques, une orchestration multi-agents et des flux de travail complexes au fur et à mesure de l'évolution des exigences.

Exemple concret de LlamaIndex

Cet exemple porte sur une filiale d'une entreprise aérospatiale spécialisée dans les solutions de navigation et d'exploitation aériennes. Ils doivent relever un défi croissant qui consiste à piloter des essais de chatbots basés sur l'IA non coordonnés. Les essais ont donné lieu à des travaux répétés, à de longs cycles de développement, à des obstacles à la conformité et à des mises en œuvre isolées au sein de l'organisation.

Ils ont développé un framework d'agents unifié, une solution réutilisable basée sur des modèles basée sur le framework LlamaIndex open source qui rend la création d'agents beaucoup plus efficace. Ils ont comparé plusieurs frameworks concurrents, à la fois orientés chaîne et basés sur des graphes. En fin de compte, ils ont opté LlamaIndex pour trois avantages essentiels : sa conception flexible, ses composants modulaires et ses commandes d'orchestration prêtes pour la production.

La plate-forme réduit le temps de développement et de déploiement des agents de 87 %, passant de 512 à 64 heures. Cette réduction a été réalisée en permettant aux équipes de créer des agents avec environ 50 lignes de code et un fichier de configuration JSON. Les équipes ont tiré parti d'un cadre unifié intégrant la sécurité, la conformité et un accès privilégié au système. Pour plus de détails, consultez les études de cas LlamaIndex clients.

Comparaison des frameworks d'IA agentiques

Lorsque vous sélectionnez un framework d'IA agentic pour le développement d'agents autonomes, réfléchissez à la manière dont chaque option correspond à vos besoins spécifiques. Tenez compte non seulement de ses capacités techniques, mais également de son adéquation organisationnelle, notamment de l'expertise de l'équipe, de l'infrastructure existante et des exigences de maintenance à long terme. De nombreuses organisations pourraient bénéficier d'une approche hybride, en tirant parti de plusieurs cadres pour les différents composants de leur écosystème d'IA autonome.

Le tableau suivant compare les niveaux de maturité (le plus fort, le plus fort, le plus adéquat ou le plus faible) de chaque framework selon les principales dimensions techniques. Pour chaque framework, le tableau inclut également des informations sur les options de déploiement en production et la complexité de la courbe d'apprentissage.

Cadre	AWS intégration	Support multi-agents autonome	Complexité du workflow autonome	Capacités multimodales	Sélection du modèle de fondation	Intégration de l'API LLM	Déploiement en production	Courbe d'apprentissage
AutoGen	Faible	Fort	Fort	Suffisant	Suffisant	Fort	Faites-le vous-même (DIY)	Raide
CrewAI	Faible	Fort	Suffisant	Faible	Suffisant	Suffisant	DIY	Modérée
LangChain / LangGraph	Suffisant	Fort	Le plus fort	Le plus fort	Le plus fort	Le plus fort	Plateforme ou bricolage	Raide
LlamaIndex	Suffisant	Suffisant	Fort	Suffisant	Fort	Fort	Plateforme ou bricolage	Modérée
Strands Agents	Le plus fort	Fort	Le plus fort	Fort	Fort	Le plus fort	DIY	Modérée

Considérations à prendre en compte lors du choix d'un framework d'IA agentic

Lorsque vous développez des agents autonomes, tenez compte des facteurs clés suivants :

- AWS intégration de l'infrastructure — Les organisations fortement investies AWS bénéficieront le plus des intégrations natives de Strands Agents with Services AWS pour les flux de travail autonomes. Pour plus d'informations, consultez le résuméAWS hebdomadaire (AWS blog).
- Sélection du modèle de base : déterminez quel framework fournit le meilleur support pour vos modèles de base préférés (par exemple, les modèles Amazon Nova sur Amazon Bedrock ou Anthropic Claude), en fonction des exigences de raisonnement de votre agent autonome. Pour plus d'informations, consultez la section Création d'agents efficaces sur le Anthropic site Web.
- Intégration de l'API LLM : évaluez les frameworks en fonction de leur intégration avec vos interfaces de service LLM (Large Language Model) préférées (par exemple, Amazon Bedrock ouOpenAI) pour le déploiement en production. Pour plus d'informations, consultez la section Model Interfaces dans la Strands Agents documentation.
- Exigences multimodales — Pour les agents autonomes qui doivent traiter du texte, des images et de la parole, tenez compte des capacités multimodales de chaque framework. Pour plus d'informations, consultez la section Multimodalité dans la LangChain documentation.
- Complexité du flux de travail autonome — Des flux de travail autonomes plus complexes dotés d'une gestion d'état sophistiquée peuvent favoriser les capacités avancées des machines à états. LangGraph
- Collaboration d'équipe autonome — Les projets qui nécessitent une collaboration autonome explicite basée sur les rôles entre des agents spécialisés peuvent bénéficier de l'architecture orientée équipe de. CrewAI
- Paradigme de développement autonome — Les équipes qui préfèrent les modèles conversationnels et asynchrones pour les agents autonomes peuvent préférer l'architecture événementielle de. AutoGen
- Approche gérée ou basée sur le code — Les organisations qui souhaitent une expérience entièrement gérée avec un minimum de codage devraient envisager Amazon Bedrock Agents. Organisations nécessitant une personnalisation plus poussée peuvent Strands Agents préférer d'autres frameworks dotés de fonctionnalités spécialisées qui répondent mieux aux exigences spécifiques des agents autonomes.

- Préparation à la production pour les systèmes autonomes : considérez les options de déploiement, les capacités de surveillance et les fonctionnalités d'entreprise pour les agents autonomes de production.

Plateformes

Les plateformes Agentic AI fournissent les couches d'exécution, d'orchestration et d'intégration de base nécessaires au déploiement, à la mise à l'échelle et à la gestion des systèmes agentic de niveau production. Les frameworks définissent la manière dont les agents sont créés et les protocoles régissent leur mode de communication. Les plateformes fournissent l'environnement dans lequel ces agents opèrent, collaborent et évoluent en toute sécurité à grande échelle.

Les plateformes Agentic combinent des fonctionnalités d'exécution de modèles, de gestion du contexte, d'intégration d'outils, d'observabilité et de gouvernance dans des environnements unifiés. Ces plateformes permettent aux entreprises de passer de l'expérimentation au déploiement à l'échelle de l'entreprise.

Dans cette section :

Pourquoi les plateformes sont importantes

Les plateformes d'IA agentic sont essentielles pour les organisations qui cherchent à opérationnaliser des systèmes autonomes en production. Ils offrent les fonctionnalités suivantes :

- Fournissez une orchestration d'exécution pour l'hébergement, le dimensionnement et la coordination des agents.
- Gérez l'état, le contexte et la mémoire dans les flux de travail multi-agents.
- Proposez des contrôles de sécurité, d'identité et de gouvernance conformes aux normes de l'entreprise.
- Intégrez les écosystèmes d'outillage et les systèmes externes par le biais de normes APIs ou de protocoles.
- Favorisez l'observabilité et l'auditabilité des interactions avec les agents et des flux d'événements.

- Support de l'interopérabilité entre modèles, permettant aux agents d'utiliser plusieurs modèles de base dans un seul environnement.

Ces fonctionnalités transforment les agents individuels en systèmes coordonnés et adaptatifs capables de fonctionner de manière fiable dans les limites de l'entreprise et des réglementations.

Types de plateformes d'IA agentique

Les plateformes d'IA agentic appartiennent généralement à une ou plusieurs des catégories suivantes :

- Agent géré : les plateformes entièrement gérées fournissent une infrastructure, une mémoire et des capacités d'orchestration intégrées. Ils réduisent les frais d'exploitation et accélèrent le délai de production.
- Orchestration open source — Les plateformes agentiques open source offrent flexibilité et transparence aux entreprises qui préfèrent des environnements personnalisables ou un déploiement sur site.
- Entreprise hybride : les plateformes hybrides intègrent des composants gérés et auto-hébergés, alliant l'évolutivité des services gérés dans le cloud au contrôle des systèmes d'entreprise.

Considérations relatives au choix des plateformes

Lors de la sélection ou de la conception d'une plateforme d'IA agentic, les organisations doivent prendre en compte les points suivants :

- Profondeur de l'intégration : évaluez dans quelle mesure la plateforme s'intègre aux sources de données, aux outils et aux protocoles existants.
- Évolutivité — Assurez-vous que la plateforme peut évoluer de manière dynamique pour prendre en charge les charges de travail autonomes et la collaboration entre plusieurs agents.
- Sécurité et conformité : évaluez les fonctionnalités de confidentialité, de chiffrement et de gouvernance des données par rapport aux exigences organisationnelles et régionales.
- Extensibilité — Choisissez des plateformes dotées d'architectures modulaires qui permettent d'ajouter de nouveaux outils, modèles ou agents au fil du temps.
- Observabilité — Préférez les plateformes qui fournissent une télémétrie détaillée, une traçabilité et des journaux d'audit pour les interactions agentiques.

- Rentabilité — Envisagez des modèles sans serveur ou basés sur l'utilisation afin d'optimiser le coût des charges de travail variables.

Agents Amazon Bedrock

Amazon Bedrock Agents est un service entièrement géré qui vous permet de créer et de configurer des agents autonomes dans vos applications. Il peut orchestrer les interactions entre les modèles de base, les sources de données, les applications logicielles et les conversations avec les utilisateurs. Son approche rationalisée de création d'agents ne vous oblige pas à fournir de la capacité, à gérer l'infrastructure ou à écrire du code personnalisé.

Principales fonctionnalités d'Amazon Bedrock Agents

Amazon Bedrock Agents inclut les fonctionnalités clés suivantes :

- Service entièrement géré : gestion complète de l'infrastructure sans qu'il soit nécessaire de fournir de la capacité ou de gérer les systèmes sous-jacents. Pour plus d'informations, consultez Automatiser les tâches de votre application à l'aide d'agents d'intelligence artificielle dans la documentation Amazon Bedrock.
- Développement piloté par API : définissez et exécutez des agents via de simples appels d'API en spécifiant des modèles, des instructions, des outils et des paramètres de configuration. Pour plus d'informations, consultez la section Créer et configurer un agent manuellement dans la documentation Amazon Bedrock.
- Groupes d'actions : définissez les actions spécifiques que votre agent peut effectuer en créant des groupes d'actions avec des schémas d'API. Pour plus d'informations, consultez la section Utiliser des groupes d'actions pour définir les actions que votre agent doit effectuer dans la documentation Amazon Bedrock.
- Intégration aux bases de connaissances — Connectez-vous facilement aux bases de connaissances Amazon Bedrock pour augmenter les réponses des agents grâce aux données de votre organisation. Pour plus d'informations, consultez Augmentez la génération de réponses pour votre agent grâce à la base de connaissances de la documentation Amazon Bedrock.
- Modèles d'invite avancés : personnalisez le comportement des agents grâce à des modèles d'invite pour le prétraitement, l'orchestration, la génération de réponses dans la base de connaissances et le post-traitement. Pour plus d'informations, consultez la section Améliorer la précision de l'agent à l'aide de modèles d'invite avancés dans Amazon Bedrock dans la documentation Amazon Bedrock.

- Traçabilité et observabilité : suivez le processus de step-by-step raisonnement de l'agent à l'aide des fonctionnalités de suivi intégrées. Pour plus d'informations, consultez la section Suivi du processus de step-by-step raisonnement de l'agent à l'aide de trace dans la documentation Amazon Bedrock.
- Gestion des versions et alias : créez plusieurs versions de votre agent et déployez-les via des alias pour des déploiements contrôlés. Pour plus d'informations, consultez la section Déployer et utiliser un agent Amazon Bedrock dans votre application dans la documentation Amazon Bedrock.

Quand utiliser Amazon Bedrock Agents

Amazon Bedrock Agents est particulièrement adapté aux scénarios d'agents autonomes, notamment :

- Organisations qui souhaitent bénéficier d'une expérience entièrement gérée pour créer et déployer des agents sans gérer l'infrastructure
- Projets nécessitant un développement et un déploiement rapides d'agents par le biais de la configuration plutôt que du code
- Cas d'utilisation bénéficiant d'une intégration étroite avec d'autres fonctionnalités d'Amazon Bedrock, telles que les bases de connaissances et les garde-corps
- Des équipes qui ne disposent pas des ressources internes nécessaires pour créer des agents à partir de zéro, mais qui ont besoin de capacités autonomes prêtes à être mises en production

Approche de mise en œuvre pour Amazon Bedrock Agents

Amazon Bedrock Agents propose une approche de mise en œuvre basée sur la configuration pour les parties prenantes de l'entreprise. Le service permet aux organisations de :

- Définissez les agents via les appels d'API AWS Management Console or sans écrire de code complexe.
- Créez des groupes d'actions qui spécifient les opérations APIs et que l'agent peut effectuer.
- Connectez les bases de connaissances pour fournir des informations spécifiques au domaine à l'agent.
- Testez et répétez le comportement des agents via une interface visuelle.

Cette approche gérée permet aux équipes commerciales de développer et de déployer rapidement des agents autonomes sans expertise technique approfondie en matière de développement de modèles d'IA ou de gestion d'infrastructure.

Exemple concret d'Amazon Bedrock Agents

Une solution d'opérations financières (FinOps) décrite dans ce billet de AWS blog utilise le framework multi-agents Amazon Bedrock pour créer un assistant de gestion des coûts dans le cloud piloté par l'IA. Le modèle économique de base d'Amazon Nova alimente la solution dans le cadre de laquelle un agent FinOps superviseur central délègue des tâches à des agents spécialisés. Ces agents récupèrent et analysent les données de AWS dépenses en utilisant AWS Cost Explorer et génèrent des recommandations de réduction des coûts en utilisant. AWS Trusted Advisor

Le système inclut un accès utilisateur sécurisé via Amazon Cognito, une interface hébergée sur AWS Amplify, et des groupes d' AWS Lambda action pour des analyses et des prévisions en temps réel. Les équipes financières peuvent poser des questions en langage naturel telles que « Quels étaient mes coûts en février 2025 ? » Le système répond par des ventilations détaillées, des suggestions d'optimisation et des prévisions, le tout dans le cadre d'une architecture évolutive et sans serveur déployée par l'utilisateur. AWS CloudFormation

Amazon Bedrock AgentCore

Amazon Bedrock AgentCore est une plateforme agentique permettant de créer, de déployer et d'exploiter des agents hautement performants en toute sécurité et à grande échelle, en utilisant n'importe quel framework, modèle ou protocole. En utilisant AgentCore, vous pouvez effectuer les opérations suivantes, le tout sans aucune gestion d'infrastructure :

- Créez des agents plus rapidement.
- Permettez aux agents de prendre des mesures sur l'ensemble des outils et des données.
- Exécutez les agents en toute sécurité avec une faible latence et des durées d'exécution prolongées.
- Surveillez les agents en production.

AgentCore élimine le fardeau indifférencié lié à la mise en place d'une infrastructure d'agents spécialisés, ce qui vous permet d'accélérer la mise en production de vos agents. Ses services peuvent être utilisés ensemble ou indépendamment et sont compatibles avec n'importe quel framework, y compris CrewAILangGraph,LlamaIndex, etStrands Agents. AgentCoreest également

compatible avec tous les modèles de fondations disponibles dans ou en dehors d'Amazon Bedrock, offrant ainsi une flexibilité ultime.

AgentCore est composé de plusieurs services clés :

- Amazon Bedrock AgentCore Runtime — Fournit un environnement sécurisé, évolutif et sans serveur pour héberger et exécuter vos agents, sans avoir à gérer l'infrastructure requise pour le déploiement et l'exécution d'agents ou d'outils d'IA.
- Amazon Bedrock AgentCore Memory : propose un système de mémoire géré, permettant aux agents de conserver le contexte des interactions pour des conversations plus personnalisées et cohérentes en conservant des connaissances à la fois immédiates et à long terme.
- Amazon Bedrock AgentCore Gateway — Simplifie le processus de création, de sécurisation et de recherche des bons outils pour les agents. Avec AgentCore Gateway, les développeurs peuvent convertir APIs les fonctions Lambda et les services existants en outils compatibles avec le Model Context Protocol (MCP) et les mettre à la disposition des agents.
- Amazon Bedrock AgentCore Identity — Fournit un service de gestion des identités et des accès des agents sécurisé et évolutif qui accélère le développement d'agents IA. Avec AgentCore Identity, vous pouvez attribuer des identités uniques et vérifiables aux agents, ce qui permet un contrôle d'accès précis et des interactions sécurisées pilotées par les agents avec les systèmes de l'entreprise.
- Outils AgentCore intégrés Amazon Bedrock : vous permet d'utiliser des outils intégrés pour améliorer votre flux de travail de développement et de test. Utilisez ces outils pour interagir efficacement avec votre application, en permettant aux agents d'intelligence artificielle d'écrire et d'exécuter du code en toute sécurité dans des environnements sandbox. Utilisez l'outil de navigation pour permettre aux agents d'intelligence artificielle d'interagir avec des sites Web à grande échelle.
- Amazon Bedrock AgentCore Observability : fournit des fonctionnalités de journalisation et de surveillance, vous donnant une visibilité en temps réel sur les performances et le comportement de votre agent afin de faciliter le débogage et l'optimisation.

Principales fonctionnalités de AgentCore

AgentCore inclut les principales fonctionnalités suivantes :

- Entièrement géré et extensible : AgentCore il s'agit d'un service entièrement géré, ce qui signifie qu'il AWS gère l'infrastructure et la maintenance sous-jacentes. Il est également extensible, ce

qui vous permet de personnaliser et d'améliorer les fonctionnalités de vos agents. Pour plus d'informations, voir Commencer avec AgentCore Runtime dans la AgentCore documentation.

- Mémoire à long terme et à court terme : offrez des interactions plus personnalisées et pertinentes en dotant les agents d'un système de mémoire capable de mémoriser le contexte des conversations en cours et des connaissances à long terme. Pour plus d'informations, voir Commencer avec AgentCore la mémoire dans la AgentCore documentation.
- Développement et intégration d'outils simplifiés : permettez à vos agents de découvrir et d'utiliser des outils via un point de terminaison unique et sécurisé. Transformez rapidement les ressources existantes de votre entreprise en outils prêts à être utilisés par les agents en quelques lignes de code, ce qui permet aux développeurs de se concentrer sur le développement de fonctionnalités uniques. Pour plus d'informations, voir Commencer avec AgentCore Gateway dans la AgentCore documentation.
- Infrastructure sécurisée et évolutive : AgentCore fournit un environnement sécurisé et évolutif pour le déploiement et l'exploitation des agents. Il inclut des fonctionnalités de gestion des identités et des accès, de chiffrement des données et de sécurité du réseau. Pour plus d'informations, voir Commencer avec AgentCore Identity dans la AgentCore documentation.
- Intégration à un large éventail d'outils : vous permet d'intégrer à vos agents une variété d'outils, notamment un interpréteur de code et un outil de navigateur que vous pouvez créer à l'aide des outils AgentCore intégrés. Pour plus d'informations, voir Commencer avec l'interpréteur de AgentCore code et Commencer avec le AgentCore navigateur dans la AgentCore documentation.
- Observabilité et surveillance complètes : bénéficiez d'une visibilité approfondie sur vos agents grâce à des outils complets permettant de suivre, de déboguer et de surveiller leurs performances en production. Visualisez le parcours d'exécution complet de l'agent pour auditer son raisonnement et résoudre les défaillances. Utilisez des tableaux de bord en temps réel et des données de télémétrie standardisées pour suivre les indicateurs opérationnels clés. Pour plus d'informations, consultez la section Ajouter de l'observabilité à vos AgentCore ressources Amazon Bedrock dans la AgentCore documentation.

Quand utiliser AgentCore

AgentCore est particulièrement bien adapté aux scénarios d'agents autonomes, notamment :

- Organisations qui souhaitent accélérer le développement et réduire les frais opérationnels grâce à un service entièrement géré qui gère l'infrastructure, la sécurité, les outils intégrés, l'observabilité et la mise à l'échelle

- Projets nécessitant de la flexibilité avec des services modulaires qui fonctionnent ensemble ou indépendamment et sont compatibles avec n'importe quel framework, similaire CrewAI ouLangGraph, et avec n'importe quel modèle de base, quelle que soit la source
- Cas d'utilisation nécessitant des agents conversationnels dynamiques qui doivent conserver le contexte et tirer des leçons des interactions passées pour fournir des réponses personnalisées et pertinentes
- Les agents sont en mesure d'effectuer des tâches complexes grâce à une intégration simple à diverses applications, sources de données et APIs

Approche de mise en œuvre pour AgentCore

AgentCore est conçu pour les organisations qui souhaitent faire passer les agents d'intelligence artificielle de la preuve de concept, construite à l'aide de frameworks d'agents open source ou personnalisés, à la production. Grâce à AgentCore ces outils, les organisations peuvent effectuer les opérations suivantes :

- Déployez des agents en toute sécurité sur une infrastructure sans serveur, compatible avec n'importe quel framework et modèle, avec isolation des sessions et gestion intégrée des identités et des accès pour garantir la end-to-end sécurité et la conformité. Créez rapidement des agents AgentCore Runtime pour les principaux frameworks d'agents à l'aide du kit de démarrage.
- Améliorez les agents en intégrant une mémoire persistante pour la rétention du contexte, en simplifiant le développement et l'intégration des outils via AgentCore Gateway. Tirez parti de l'outil de navigateur intégré et de l'interpréteur de code pour des flux de travail avancés.
- Suivez, déboguez et surveillez les agents d'IA en production à l'aide de tableaux de bord d'observabilité optimisés par Amazon CloudWatch Application Insights et OpenTelemetry en suivant les indicateurs clés des AgentCore ressources (temps d'exécution, mémoire, passerelle et outils).
- Accélérez le déploiement et l'innovation avec des services modulaires entièrement gérés, des blocs composables ensemble ou indépendamment, avec n'importe quel framework d'agents et fournisseur de modèles. Cette flexibilité permet aux entreprises de passer plus rapidement du prototype à la production.

Cette approche gérée permet aux entreprises de créer, déployer et exécuter rapidement et en toute sécurité des agents d'intelligence artificielle et des systèmes multi-agents de qualité professionnelle à n'importe quelle échelle.

Exemple concret de AgentCore

AWS a observé que l'une des plus grandes banques d'Amérique latine propose AI/ML depuis des années une expérience bancaire numérique hyperpersonnalisée et sécurisée. La banque développe les services d'intelligence artificielle agentic en les utilisant AgentCore pour fournir aux clients des interactions intuitives, une sécurité renforcée et une automatisation accrue. Selon le CTO, AgentCore il devrait soutenir leurs efforts pour respecter les engagements des clients à grande échelle. AgentCore fournit à leurs développeurs les outils et la flexibilité nécessaires pour créer et gérer des agents, tout en garantissant le respect des réglementations financières.

Protocoles

Les agents d'intelligence artificielle ont besoin de protocoles de communication standardisés pour interagir avec les autres agents et services. Organisations qui mettent en œuvre des architectures d'agents sont confrontées à des défis importants en matière d'interopérabilité, d'indépendance des fournisseurs et de pérennisation de leurs investissements.

Cette section vous aide à naviguer dans le paysage des agent-to-agent protocoles en mettant l'accent sur les normes ouvertes qui maximisent la flexibilité et l'interopérabilité. (Pour plus d'informations sur agent-to-tool les protocoles, voir Stratégie d'intégration des outils plus loin dans ce guide.)

Cette section met en évidence le Model Context Protocol (MCP), un standard ouvert initialement développé Anthropic en 2024. Aujourd'hui, soutient AWS activement le MCP en contribuant au développement et à la mise en œuvre du protocole. AWS collabore avec les principaux frameworks d'agents open source, notammentLangGraph, et CrewAILlamaIndex, pour façonner le futur de la communication inter-agents sur le protocole. Pour plus d'informations, voir Protocoles ouverts pour l'interopérabilité des agents, partie 1 : Communication entre agents sur MCP (AWS blog).

Dans cette section :

- Pourquoi le choix du protocole est important
- Agent-to-agent protocoles
- Sélection de protocoles agentiques
- Stratégie de mise en œuvre des protocoles agentiques
- Commencer à utiliser MCP
- ???

Pourquoi le choix du protocole est important

La sélection du protocole façonne fondamentalement la manière dont vous pouvez créer et faire évoluer votre architecture d'agent d'IA. En choisissant des protocoles garantissant la portabilité entre les frameworks d'agents, vous bénéficiez de la flexibilité nécessaire pour combiner différents systèmes d'agents et flux de travail pour répondre à vos besoins spécifiques.

Les protocoles ouverts vous permettent d'intégrer des agents dans plusieurs frameworks. Par exemple, utilisez-le LangChain pour le prototypage rapide et implémentez des systèmes de production communiquant via un protocole commun, tel que MCP ou le protocole Agent2Agent (A2A). Strands Agents Cette flexibilité réduit la dépendance à l'égard de fournisseurs d'IA spécifiques, simplifie l'intégration aux systèmes existants et vous permet d'améliorer les capacités des agents au fil du temps.

Des protocoles bien conçus établissent également des modèles de sécurité cohérents pour l'authentification et l'autorisation au sein de votre écosystème d'agents. Plus important encore, la portabilité des protocoles préserve votre liberté d'adopter de nouveaux frameworks et fonctionnalités d'agents au fur et à mesure de leur apparition. Le choix de protocoles ouverts protège votre investissement dans le développement d'agents tout en maintenant l'interopérabilité avec les systèmes tiers.

Avantages des protocoles ouverts

Lorsque vous implémentez vos propres extensions ou que vous créez des systèmes d'agents personnalisés, les protocoles ouverts offrent des avantages indéniables :

- Documentation et transparence — Fournissez généralement une documentation complète et des implémentations transparentes
- Support communautaire : accès à des communautés de développeurs plus larges pour le dépannage et les meilleures pratiques
- Garanties d'interopérabilité : meilleure assurance que vos extensions fonctionneront sur différentes implémentations
- Compatibilité future : réduction du risque d'interruption des modifications ou de dépréciation
- Influence sur le développement — Possibilité de contribuer à l'évolution du protocole

Agent-to-agent protocoles

Le tableau suivant fournit une vue d'ensemble des protocoles agentiques qui permettent à plusieurs agents de collaborer, de déléguer des tâches et de partager des informations.

Protocole	Idéal pour	Considérations

Communication entre agents MCP	Organisations à la recherche de modèles de collaboration flexibles entre agents	• Une extension du protocole MCP (Model Context Protocol) proposée par AWS qui s'appuie sur ses bases de agent-to-agent communication existantes • Permet une collaboration fluide entre les agents grâce OAuth à une sécurité basée
Protocole A2A	Écosystèmes d'agents multiplateformes	• Soutenu par Google • Norme plus récente avec une adoption plus limitée par rapport au MCP

Choix des options de protocole

Lors de agent-to-agent la mise en œuvre de la communication, adaptez vos exigences de communication spécifiques aux capacités de protocole appropriées. Les différents modèles d'interaction nécessitent des fonctionnalités de protocole différentes. Le tableau suivant décrit les modèles de communication courants et recommande les choix de protocole les plus adaptés à chaque scénario.

Modèle	Description	Choix de protocole idéal
Demande et réponse simples	Interactions ponctuelles entre agents	MCP avec flux apatrides
Dialogues dynamiques	Conversations continues avec le contexte	MCP avec gestion de session
Collaboration multi-agents	Interactions complexes entre plusieurs agents	Inter-agent MCP ou AutoGen
Flux de travail basés sur l'équipe	Des équipes d'agents hiérarchiques avec des rôles définis	Inter-agent MCP, ou CrewAI AutoGen

Au-delà des modèles de communication, plusieurs facteurs techniques et organisationnels peuvent influencer le choix de votre protocole. Le tableau suivant décrit les principales considérations qui peuvent vous aider à déterminer quel protocole correspond le mieux à vos exigences de mise en œuvre spécifiques.

Considération	Description	Exemple
Modèle de sécurité	Exigences en matière d'authentification et d'autorisation	OAuth 2,0 en MCP
Environnement de déploiement	Où les agents courront et communiqueront	Machine distribuée ou unique
Compatibilité avec les écosystèmes	Intégration avec les frameworks d'agents existants	LangChain ou Strands Agents
Besoins d'évolutivité	Croissance attendue des interactions entre agents	Capacités de streaming de MCP

Sélection de protocoles agentiques

Pour la plupart des organisations qui créent des systèmes d'agents de production, le protocole MCP (Model Context Protocol) constitue la base de communication la plus complète et la mieux prise en charge. agent-to-agent MCP bénéficie des contributions de développement actives de la part de la communauté open source AWS et de celle-ci.

Il est important de sélectionner les bons protocoles agentic pour les organisations qui cherchent à mettre en œuvre efficacement l'IA agentic. Les considérations varient en fonction du contexte organisationnel.

Considérations relatives à la sélection du protocole agentic

Organisations devraient prendre en compte les meilleures pratiques suivantes lors de la sélection de protocoles pour les systèmes d'IA agentic :

- Prioriser les standards ouverts — Les organisations devraient adopter des protocoles ouverts tels que le MCP pour garantir l'interopérabilité et l'extensibilité à long terme et pour réduire le risque de dépendance vis-à-vis d'un fournisseur.
- Équilibre entre rapidité et flexibilité — Les entreprises en démarrage et les premiers utilisateurs peuvent commencer par utiliser des protocoles propriétaires bien pris en charge pour un développement rapide, mais doivent définir une voie de migration vers des standards ouverts à mesure que les systèmes arrivent à maturité.
- Implémentation de couches d'abstraction : les entreprises doivent mettre en œuvre l'abstraction des protocoles pour simplifier la migration, permettre l'adoption hybride et mettre en œuvre des stratégies d'intégration pérennes.
- Mettre l'accent sur la sécurité et la conformité — Les organisations des secteurs réglementés doivent sélectionner des protocoles dotés de fonctionnalités d'authentification, de chiffrement et d'audit robustes pour répondre aux exigences de gouvernance et de conformité.
- Évaluer la maturité de l'écosystème — Toutes les organisations devraient évaluer l'état de santé, l'adoption et le soutien communautaire de chaque protocole afin de garantir la durabilité et de minimiser la dette technique.
- S'engager dans l'élaboration de normes — Les organisations devraient participer à des organismes de normalisation ou à des communautés open source pour contribuer à façonner l'évolution des protocoles et influencer les meilleures pratiques.
- Tenez compte de la souveraineté des données — Le gouvernement et les secteurs réglementés doivent veiller à ce que les choix de protocoles soient conformes aux exigences de résidence et de souveraineté des données dans les régions de déploiement.
- Tirez parti des services gérés : dans la mesure du possible, utilisez des implémentations gérées ou sans serveur de protocoles agentic pour réduire la complexité opérationnelle et accélérer le déploiement.

Stratégie de mise en œuvre des protocoles agentiques

Pour mettre en œuvre efficacement les protocoles agentic au sein de votre organisation, considérez les étapes stratégiques suivantes :

1. Commencez par l'alignement des normes — Adoptez des protocoles ouverts établis dans la mesure du possible.

2. Créez des couches d'abstraction : implémentez des adaptateurs entre vos systèmes et des protocoles spécifiques.
3. Contribuez aux standards ouverts — Participez aux communautés de développement de protocoles.
4. Surveillez l'évolution des protocoles : restez informé des nouvelles normes et des mises à jour.
5. Testez régulièrement l'interopérabilité : vérifiez que vos implémentations restent compatibles.

Commencer à utiliser MCP

AWS soutient activement le Model Context Protocol (MCP) en contribuant au développement et à la mise en œuvre du protocole. AWS collabore avec les principaux frameworks d'agents open source, notammentLangGraph, et CrewAILlamaIndex, pour façonner le futur de la communication inter-agents sur le protocole.

Pour implémenter le MCP dans l'architecture de votre agent, effectuez les actions suivantes :

1. Explorez les implémentations de MCP dans des frameworks tels que le Strands Agents SDK.
2. Consultez la documentation technique du Model Context Protocol.
3. Lisez Protocoles ouverts pour l'interopérabilité des agents, partie 1 : communication entre agents sur MCP (AWS blog) pour en savoir plus sur l'interopérabilité des agents.
4. Rejoignez la communauté MCP pour influencer l'évolution du protocole.

Le MCP fournit une couche de communication qui permet aux agents d'interagir avec des données et des services externes et peut également être utilisée pour permettre aux agents d'interagir avec d'autres agents. L'implémentation du transport HTTP Streamable du protocole fournit aux développeurs un ensemble complet de modèles d'interaction sans avoir à réinventer la roue. Ces modèles prennent en charge à la fois les request/response flux apatrides et la gestion dynamique des sessions avec persistance. IDs

En adoptant des protocoles ouverts tels que le MCP, vous positionnez votre organisation pour créer des systèmes d'agents qui restent flexibles, interopérables et adaptables à mesure que la technologie de l'IA évolue. Pour plus d'informations sur la mise en œuvre agent-to-tool du protocole, voir Stratégie d'intégration des outils plus loin dans ce guide.

Commencer à utiliser A2A

Le protocole Agent2Agent (A2A) permet une collaboration décentralisée entre les agents via une couche sémantique partagée. Au lieu d'acheminer tout le travail via un orchestrateur central, A2A permet aux agents de se découvrir, de promouvoir leurs capacités, de négocier des tâches et de partager le contexte à l'aide d'un protocole léger basé sur JSON. Chaque agent publie un manifeste de capacités.

L'exemple suivant montre un manifeste de fonctionnalités A2A simplifié qui annonce les actions prises en charge par un agent, les entrées requises et les métadonnées opérationnelles pour permettre la découverte et la négociation des tâches :

```
{
  "can": ["summarize.text", "extract.keywords"],
  "needs": ["document.input"],
  "meta": { "version": "1.0.3", "latencyMs": 120 }
}
```

Ce modèle permet l'appariement dynamique des capacités, la délégation en milieu de tâche et la collaboration interorganisationnelle. Les agents peuvent s'auto-organiser autour des tâches, former des groupes de travail temporaires et s'adapter à l'entrée ou à la sortie de nouvelles fonctionnalités dans le système.

A2A prend en charge des interactions allant de simples demandes apatrides à des sessions de négociation en plusieurs étapes, notamment :

- peer-to-peerMessagerie directe pour une collaboration à faible latence
- Négociation sémantique des tâches, où les agents sélectionnent le pair le plus approprié
- Découverte basée sur les capacités, permettant une division émergente du travail
- Ancrage de session pour des interactions dynamiques en plusieurs étapes

En adoptant des protocoles ouverts natifs pour les agents tels que l'A2A, les entreprises créent des systèmes d'IA modulaires, interopérables et capables de collaboration transfrontalière. L'A2A garantit que les écosystèmes d'agents restent flexibles et peuvent évoluer à mesure que de nouveaux agents, équipes ou systèmes externes sont introduits, sans nécessiter de couches d'orchestration rigides ni de couplage préalable.

Pour implémenter le protocole A2A dans l'architecture de votre agent, effectuez les actions suivantes :

1. Consultez la spécification du protocole A2A — Lisez la dernière version de la spécification du protocole Agent2Agent (A2A) pour savoir comment fonctionnent les manifestes de capacités, les flux de négociation et la poignée de main des agents.
2. Explorez les environnements d'exécution compatibles avec A2A : évaluez les frameworks tels que le SDK Strands Agents ou les couches d'exécution personnalisées qui prennent en charge les manifestes de capacités et la négociation de style A2A. peer-to-peer
3. Implémentez un manifeste de capacités pour vos agents : définissez les `meta` champs et les champs de `can` chaque agent pour permettre la découverte, le matchmaking et la collaboration au niveau de l'intention. `needs`
4. Testez les modèles de négociation A2A : utilisez la boucle demande-offre-acceptation, les requêtes de capacité structurées ou la découverte basée sur des ragots pour comprendre comment les agents réfléchissent à qui doit gérer une tâche.
5. Testez l'A2A dans un environnement d'infrastructure mixte : associez la négociation A2A entre pairs à un routage d'événements natif via AWS Amazon EventBridge pour évaluer les modèles de coordination hybrides.
6. Rejoignez la communauté A2A : participez au groupe de travail ouvert pour vous tenir au courant des extensions, des recommandations de sécurité et des améliorations de l'interopérabilité entre fournisseurs, et contribuer au développement du protocole.

Outils

Les agents d'intelligence artificielle apportent de la valeur en interagissant avec des APIs outils externes et des sources de données pour effectuer des tâches utiles. La bonne stratégie d'intégration des outils a un impact direct sur les capacités de votre agent, son niveau de sécurité et sa flexibilité à long terme.

Cette section vous aide à naviguer dans le paysage de l'intégration des outils en mettant l'accent sur les normes ouvertes qui maximisent votre liberté et votre flexibilité. La section met en évidence le protocole MCP (Model Context Protocol) pour l'intégration des outils et passe en revue les outils spécifiques au framework et les méta-outils spécialisés qui améliorent les flux de travail des agents.

Dans cette section :

Catégories d'outils

Les systèmes d'agents de construction impliquent trois catégories principales d'outils.

Outils basés sur des protocoles

Les outils basés sur des protocoles utilisent des protocoles normalisés pour la agent-to-tool communication :

- Outils MCP : outils standard ouverts qui fonctionnent dans tous les frameworks avec des options d'exécution locales et distantes.
- OpenAIappel de fonctions — Outils propriétaires spécifiques aux OpenAI modèles.
- Anthropicoutils — Variation d'une OpenAI fonction faisant appel à des outils propriétaires spécifiques aux modèles de Anthropic Claude.

Outils natifs du framework

Les outils natifs du framework sont intégrés directement dans des frameworks d'agents spécifiques :

- Strands Agents outils — quick-to-implement Outils légers, spécifiques au Strands Agents framework.
- LangChainoutils : des outils Python basés sur des outils étroitement intégrés à l'LangChainécosystème.
- LlamaIndexoutils — Outils optimisés pour la récupération et le traitement des données au sein LlamaIndex de.

Méta-outils

Les méta-outils améliorent les flux de travail des agents sans effectuer directement d'actions externes :

- Outils de flux de travail : gérez le flux d'exécution des agents, la logique de branchement et la gestion des états.
- Outils de création de graphes d'agents : coordonnez plusieurs agents dans des flux de travail complexes.
- Outils de mémoire : fournissent un stockage permanent et une récupération d'informations entre les sessions des agents.
- Outils de réflexion : permettez aux agents d'analyser et d'améliorer leurs propres performances.

Outils basés sur des protocoles

En ce qui concerne les outils basés sur des protocoles, le Model Context Protocol (MCP) fournit la base la plus complète et la plus flexible pour l'intégration des outils. Comme indiqué dans le billet de blog AWS Open Source sur l'interopérabilité des agents, AWS a adopté le protocole MCP en tant que protocole stratégique, contribuant ainsi activement à son développement.

Le tableau suivant décrit les options de déploiement de l'outil MCP.

Modèle de déploiement	Description	Idéal pour	Mise en œuvre

Basé sur un studio local	Les outils s'exécutent selon le même processus que l'agent	Développement, tests et outils simples	Rapide à mettre en œuvre sans surcharge réseau
Basé sur des événements envoyés par le serveur local (SSE)	Les outils s'exécutent localement mais communiquent via HTTP	Outils locaux plus complexes avec séparation des préoccupations	Meilleure isolation mais faible latence
Diffusable via HTTP à distance	Outils exécutés sur des serveurs distants	Environnements de production et outils partagés	Évolutif et géré de manière centralisée

Les MCP officiels SDKs sont disponibles pour créer des outils MCP :

- PythonSDK — Implémentation complète avec prise en charge complète des protocoles
- TypeScriptSDK —JavaScript/TypeScriptimplémentation pour les applications Web
- JavaSDK — Implémentation de Java pour les applications d'entreprise

Ils SDKs fournissent les éléments de base pour créer des outils compatibles MCP dans votre langage préféré, avec des implémentations cohérentes de la spécification du protocole.

En outre, AWS a implémenté le MCP dans le Strands AgentsSDK. Le Strands Agents SDK fournit un moyen simple de créer et d'utiliser des outils compatibles avec MCP. Une documentation complète est disponible dans le Strands Agents GitHub référentiel. Pour des cas d'utilisation plus simples ou lorsque vous travaillez en dehors du Strands Agents cadre, les MCP officiels SDKs proposent des implémentations directes du protocole dans plusieurs langues.

Caractéristiques de sécurité des outils MCP

Les fonctionnalités de sécurité des outils MCP sont les suivantes :

- OAuth Authentification 2.0/2.1 — Authentification conforme aux normes du secteur
- Étendue des autorisations : contrôle d'accès précis pour les outils
- Découverte des capacités des outils — Découverte dynamique des outils disponibles
- Gestion structurée des erreurs — Modèles d'erreur cohérents

Commencer à utiliser les outils MCP

Pour implémenter le MCP pour l'intégration des outils, effectuez les actions suivantes :

1. Explorez le Strands AgentsSDK pour une implémentation MCP prête pour la production.
2. Consultez la documentation technique du MCP pour comprendre les concepts de base.
3. Utilisez les exemples pratiques décrits dans ce billet de blog AWS Open Source.
4. Commencez par de simples outils locaux avant de passer aux outils distants.
5. Rejoignez la communauté MCP pour influencer l'évolution du protocole.

Découvrez AgentCore Gateway

Amazon Bedrock AgentCore Gateway fournit aux développeurs un moyen simple et sécurisé de créer, déployer, découvrir et se connecter à des outils MCP et à d'autres points de terminaison cibles à grande échelle. Avec AgentCore Gateway, les développeurs peuvent convertir APIs les AWS Lambda fonctions et les services existants en outils compatibles avec MCP. Ensuite, avec seulement quelques lignes de code, ils peuvent mettre ces outils à la disposition des agents via les points de terminaison AgentCore Gateway. AgentCore Gateway prend en charge OpenAPISmithy, et Lambda en tant que types d'entrée, et constitue la seule solution qui fournit à la fois une authentification d'entrée et une authentification de sortie complètes dans un service entièrement géré.

Outils natifs du framework

Bien que le protocole MCP (Model Context Protocol) constitue la base la plus flexible, les outils natifs du framework offrent des avantages pour des cas d'utilisation spécifiques.

Le Strands AgentsSDK propose des outils Python basés sur des outils qui se caractérisent par leur conception légère qui nécessite une surcharge minimale pour des opérations simples. Ils permettent une mise en œuvre rapide et permettent aux développeurs de créer des outils avec seulement quelques lignes de code. De plus, ils sont étroitement intégrés pour fonctionner parfaitement dans le Strands Agents cadre.

L'exemple suivant montre comment créer un outil météo simple à l'aide deStrands Agents. Les développeurs peuvent rapidement transformer les Python fonctions en outils accessibles aux agents avec une surcharge de code minimale et générer automatiquement la documentation appropriée à partir de la docstring de la fonction.

```
#Example of a simple Strands native tool

@tool

def weather(location: str) -> str:

"""Get the current weather for a location""" #

Implementation here

return f"The weather in {location} is sunny."
```

Pour le prototypage rapide ou les cas d'utilisation simples, les outils natifs du framework peuvent accélérer le développement. Toutefois, pour les systèmes de production, les outils MCP offrent une meilleure interopérabilité et une flexibilité future par rapport aux outils natifs du framework.

Le tableau suivant fournit une vue d'ensemble des autres outils spécifiques au framework.

Cadre	Type d'outil	Avantages	Considérations
AutoGen	Définitions des fonctions	Support multi-agents puissant	Microsoftécosystème
LangChain	Pythoncours	Vaste écosystème d'outils prédéfinis	Verrouillage du cadre
LlamaIndex	Fonctions Python	Optimisé pour les opérations de données	Limité à LlamaIndex

Méta-outils

Les méta-outils n'interagissent pas directement avec les systèmes externes. Ils améliorent plutôt les capacités des agents en mettant en œuvre des modèles agentiques. Cette section traite du flux de travail, du graphe de l'agent et des méta-outils de mémoire.

Méta-outils de flux de travail

Les méta-outils du flux de travail gèrent le flux d'exécution des agents :

- Gestion des états — Maintien du contexte lors des interactions entre plusieurs agents
- Logique de branchement — Activer les chemins d'exécution conditionnels
- Mécanismes de nouvelle tentative — Gérez les échecs grâce à des stratégies de relance sophistiquées

Les exemples de frameworks dotés de méta-outils de flux de travail incluent des fonctionnalités de Strands Agentsflux LangGraphde travail.

Méta-outils Agent Graph

Les méta-outils Agent Graph coordonnent la collaboration de plusieurs agents :

- Délégation de tâches — Attribuer des sous-tâches à des agents spécialisés
- Agrégation des résultats — Combinez les sorties de plusieurs agents
- Résolution des conflits — Résoudre les désaccords entre les agents

Les frameworks aiment AutoGenet CrewAIse spécialisent dans la coordination des graphes d'agents.

Méta-outils de mémoire

Les méta-outils de mémoire fournissent un stockage et une récupération persistants :

- Historique des conversations : maintenez le contexte entre les sessions
- Bases de connaissances — Stockez et récupérez des informations spécifiques au domaine
- Magasins vectoriels — Activez les fonctionnalités de recherche sémantique

Le système de ressources de MCP fournit un moyen standardisé d'implémenter des méta-outils de mémoire qui fonctionnent avec différents frameworks d'agents.

Stratégie d'intégration des outils

Le choix de la stratégie d'intégration des outils a un impact direct sur ce que vos agents peuvent accomplir et sur la facilité avec laquelle votre système peut évoluer. Priorisez les protocoles ouverts tels que le Model Context Protocol (MCP) tout en utilisant stratégiquement des outils et des méta-outils natifs du framework. Ainsi, vous pouvez créer un écosystème d'outils qui reste flexible et puissant à mesure que la technologie de l'IA progresse.

L'approche stratégique suivante en matière d'intégration des outils maximise la flexibilité tout en répondant aux besoins immédiats de votre organisation :

1. Adoptez MCP comme base : MCP fournit un moyen standardisé de connecter les agents à des outils dotés de fonctionnalités de sécurité puissantes. Commencez par utiliser MCP comme protocole d'outil principal pour :
   - Des outils stratégiques qui seront utilisés dans le cadre de la mise en œuvre de plusieurs agents.
   - Outils sensibles à la sécurité qui nécessitent une authentification et une autorisation robustes.
   - Outils nécessitant une exécution à distance dans les environnements de production.
2. Utilisez des outils natifs du framework le cas échéant — Envisagez des outils natifs du framework pour :
   - Prototypage rapide lors du développement initial.
   - Des outils simples et non critiques avec des exigences de sécurité minimales.
   - Fonctionnalité spécifique au framework qui tire parti de capacités uniques.
3. Implémentez des méta-outils pour des flux de travail complexes — Ajoutez des méta-outils pour améliorer l'architecture de vos agents :
   - Commencez simplement avec les modèles de flux de travail de base.
   - Ajoutez de la complexité à mesure que vos cas d'utilisation évoluent.
   - Standardisez les interfaces entre les agents et les méta-outils.
4. Planifiez l'évolution — Construisez en gardant à l'esprit la flexibilité future :
   - Documentez les interfaces des outils indépendamment des implémentations.
   - Créez des couches d'abstraction entre les agents et les outils.
   - Établissez des voies de migration entre des protocoles propriétaires et des protocoles ouverts.

Bonnes pratiques de sécurité pour l'intégration des outils

L'intégration des outils a un impact direct sur votre niveau de sécurité. Cette section décrit les meilleures pratiques à prendre en compte pour votre organisation.

Authentification et autorisation

Utilisez les contrôles d'accès robustes suivants :

- Utiliser OAuth 2.0/2.1 — Implémenter une authentification conforme aux normes du secteur pour les outils distants.
- Implémenter le moindre privilège : accordez aux outils uniquement les autorisations dont ils ont besoin.
- Rotation des informations d'identification : mettez régulièrement à jour les clés d'API et les jetons d'accès.

Protection des données

Pour aider à protéger les données, adoptez les mesures suivantes :

- Valider les entrées et les sorties : implémentez la validation du schéma pour toutes les interactions avec les outils.
- Chiffrez les données sensibles : utilisez le protocole TLS pour toutes les communications avec les outils distants.
- Implémenter la minimisation des données : ne transmettez que les informations nécessaires aux outils.

Surveillance et audit

Maintenez la visibilité et le contrôle en utilisant les mécanismes suivants :

- Enregistrez toutes les invocations d'outils : maintenez des pistes d'audit complètes.
- Surveillez les anomalies : détectez les modèles d'utilisation inhabituels des outils.
- Implémentez la limitation du débit : empêchez les abus liés à des appels d'outils excessifs.

Le modèle de sécurité MCP (Model Context Protocol) répond à ces préoccupations de manière globale. Pour plus d'informations, consultez la section Considérations relatives à la sécurité dans la documentation MCP.

Conclusion

Le paysage de l'IA agentique continue d'évoluer rapidement, offrant aux entreprises de nouveaux moyens puissants de créer des systèmes intelligents et autonomes. Ce guide a exploré trois éléments essentiels pour une mise en œuvre réussie : les cadres qui fournissent les bases, les plateformes qui fournissent l'environnement, les protocoles qui permettent la communication et les outils qui étendent les capacités.

À mesure que les frameworks arrivent à maturité, vous pouvez vous attendre à une interopérabilité accrue, à une standardisation autour de protocoles tels que le Model Context Protocol (MCP) et à des capacités d'orchestration plus sophistiquées pour les agents autonomes. Organisations ayant acquis une expertise avec ces frameworks aujourd'hui seront bien placées pour créer des agents de plus en plus autonomes et intelligents offrant une valeur commerciale significative.

Les plateformes fournissent l'environnement d'exécution, de gouvernance et de cycle de vie dans lequel les systèmes agentic fonctionnent. Ils répondent à des préoccupations telles que l'identité, les limites de sécurité, l'observabilité, la gestion de la mémoire, l'ancrage des sessions et l'interaction sécurisée avec les outils et les données. Dans AWS les environnements, les plateformes telles que les environnements d'exécution des agents gérés et les services d'orchestration permettent aux entreprises de déployer, de surveiller, de faire évoluer et de gouverner des agents autonomes et des systèmes agentiques à grande échelle. Les plateformes relient les cadres fondamentaux aux exigences opérationnelles réelles.

Le choix des protocoles d'agent représente une décision stratégique qui équilibre les besoins de développement immédiats avec la flexibilité et l'interopérabilité à long terme. En donnant la priorité aux protocoles ouverts et en créant des couches d'abstraction appropriées, les entreprises peuvent créer des systèmes d'agents qui restent adaptables à l'évolution des technologies tout en répondant aux exigences commerciales actuelles.

Pour la plupart des organisations, MCP représente une base solide en raison de son standard ouvert, de son écosystème en pleine croissance, de sa prise en charge des modèles de agent-to-agent communication et de ses capacités d'intégration d'outils. AWS a adopté MCP et Agent2Agent (A2A) en tant que protocoles stratégiques, contribuant activement à leur développement et à leur mise en œuvre dans des services tels que le SDK. Strands Agents En utilisant MCP ou A2A avec des outils et des méta-outils natifs du framework appropriés, vous pouvez créer des systèmes d'agents qui offrent une valeur immédiate tout en restant adaptables aux innovations futures.

Ressources

Utilisez les ressources suivantes AWS et d'autres ressources liées au développement d'agents autonomes.

AWS Blogues

- Amazon Bedrock AgentCore Memory : création d'agents sensibles au contexte
- Meilleures pratiques pour créer des applications d'IA générative robustes avec Amazon Bedrock Agents — Partie 1
- Meilleures pratiques pour créer des applications d'IA générative robustes avec Amazon Bedrock Agents — Partie 2
- Construisez de puissants pipelines RAG avec Amazon LlamaIndex Bedrock
- Créez des agents d'IA fiables avec Amazon Bedrock Observability AgentCore
- Évaluez les réponses RAG avec Amazon, Bedrock et RAGAS LlamaIndex
- Présentation de l'interpréteur de AgentCore code Amazon Bedrock
- Présentation d'Amazon Bedrock AgentCore Gateway : transformer le développement d'outils d'agent d'intelligence artificielle pour les entreprises
- Présentation d'Amazon Bedrock AgentCore Identity : sécuriser l'IA agentique à grande échelle
- Présentation Strands Agents d'un SDK Open Source pour les agents AI
- Protocoles ouverts pour l'interopérabilité des agents, partie 1 : communication entre agents sur MCP
- Lancez et faites évoluer vos agents et outils en toute sécurité sur Amazon Bedrock Runtime AgentCore
- AWS Transform pour .NET, le premier service d'intelligence artificielle agentique permettant de moderniser les applications .NET à grande échelle
- AWS Tour d'horizon hebdomadaire : Strands Agents

AWS Directives prescriptives

- Opérationnaliser l'IA agentique sur AWS
- Les fondements de l'IA agentic sur AWS

- Modèles et flux de travail 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
- Sécurité pour l'IA agentic sur AWS
- Récupérez les options et architectures de génération augmentée sur AWS

AWS ressources

- Documentation Amazon Bedrock
- Documentation Amazon Bedrock AgentCore
- Boîte à outils de AgentCore démarrage Amazon Bedrock (GitHub référentiel)
- Documentation Amazon Nova
- AWS Serveurs MCP (GitHubréférentiel)

Autres ressources

- AutoGendocumentation (Microsoft)
- Création d'agents efficaces (Anthropic)
- CrewAI GitHubréférentiel
- Documentation LangChain
- LangGraphplateforme
- Documentation LlamaIndex
- Documentation du protocole Model Context
- Documentation Strands Agents
- Strands AgentsVue d'ensemble des outils
- Strands AgentsGuide de démarrage rapide

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
Nouvelle section	Section « Plateformes » ajoutée	16 janvier 2026
Publication initiale	—	14 juillet 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 :

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

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

A

ABAC

Voir contrôle d'accès basé sur les attributs.

services abstraits

Consultez la section Services gérés.

ACIDE

Voir atomicité, consistance, isolation, durabilité.

migration active-active

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

migration active-passive

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

fonction d'agrégation

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

AI

Voir intelligence artificielle.

AIOps

Voir les opérations d'intelligence artificielle.

anonymisation

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

anti-motif

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

contrôle des applications

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

portefeuille d'applications

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

intelligence artificielle (IA)

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

opérations d'intelligence artificielle (AIOps)

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

chiffrement asymétrique

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

atomicité, cohérence, isolement, durabilité (ACID)

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

contrôle d'accès par attributs (ABAC)

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

source de données faisant autorité

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

Zone de disponibilité

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

AWS Cadre d'adoption du cloud (AWS CAF)

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

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

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

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

B

mauvais bot

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

BCP

Consultez la section Planification de la continuité des activités.

graphique de comportement

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

système de poids fort

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

classification binaire

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

filtre de Bloom

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

déploiement bleu/vert

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

bot

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

botnet

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

branche

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

accès par brise-vitre

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

stratégie existante (brownfield)

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

cache de tampon

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

capacité métier

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

planification de la continuité des activités (BCP)

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

C

CAF

Voir le cadre d'adoption du AWS cloud.

déploiement de Canary

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

CCo E

Voir le Centre d'excellence du cloud.

CDC

Voir capture des données de modification.

capture des données de modification (CDC)

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

ingénierie du chaos

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

CI/CD

Découvrez l'intégration continue et la livraison continue.

classification

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

chiffrement côté client

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

Centre d'excellence du cloud (CCoE)

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

cloud computing

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

modèle d'exploitation du cloud

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

étapes d'adoption du cloud

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

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

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

CMDB

Consultez la base de données de gestion des configurations.

référentiel de code

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

cache passif

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

données gelées

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

vision par ordinateur (CV)

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

dérive de configuration

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

base de données de gestion des configurations (CMDB)

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

pack de conformité

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

intégration continue et livraison continue (CI/CD)

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

CV

Voir vision par ordinateur.

D

données au repos

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

classification des données

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

dérive des données

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

données en transit

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

maillage de données

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

minimisation des données

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

périmètre de données

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

prétraitement des données

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

provenance des données

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

sujet des données

Personne dont les données sont collectées et traitées.

entrepôt des données

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

langage de définition de base de données (DDL)

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

langage de manipulation de base de données (DML)

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

DDL

Voir langage de définition de base de données.

ensemble profond

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

deep learning

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

defense-in-depth

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

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

administrateur délégué

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

déploiement

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

environnement de développement

Voir environnement.

contrôle de détection

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

cartographie de la chaîne de valeur du développement (DVSM)

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

jumeau numérique

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

tableau des dimensions

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

catastrophe

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

reprise après sinistre (DR)

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

DML

Voir langage de manipulation de base de données.

conception axée sur le domaine

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

DR

Voir reprise après sinistre.

détection de dérive

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

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

DVSM

Voir la cartographie de la chaîne de valeur du développement.

E

EDA

Voir analyse exploratoire des données.

EDI

Voir échange de données informatisé.

informatique de périphérie

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

échange de données informatisé (EDI)

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

chiffrement

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

clé de chiffrement

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

endianisme

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

point de terminaison

Voir point de terminaison de service.

service de point de terminaison

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

planification des ressources d'entreprise (ERP)

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

chiffrement d'enveloppe

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

environment

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

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

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

époPée

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

ERP

Voir Planification des ressources d'entreprise.

analyse exploratoire des données (EDA)

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

F

tableau des faits

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

échouer rapidement

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

limite d'isolation des défauts

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

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

branche de fonctionnalités

Voir succursale.

fonctionnalités

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

importance des fonctionnalités

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

transformation de fonctionnalité

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

invitation en quelques coups

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

FGAC

Découvrez le contrôle d'accès détaillé.

contrôle d'accès détaillé (FGAC)

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

migration instantanée (flash-cut)

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

FM

Voir le modèle de fondation.

modèle de fondation (FM)

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

G

IA générative

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

blocage géographique

Voir les restrictions géographiques.

restrictions géographiques (blocage géographique)

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

Flux de travail Gitflow

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

image dorée

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

stratégie inédite

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

barrière de protection

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

H

HA

Découvrez la haute disponibilité.

migration de base de données hétérogène

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

haute disponibilité (HA)

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

modernisation des historiens

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

données de rétention

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

migration de base de données homogène

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

données chaudes

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

correctif

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

période de soins intensifs

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

I

IaC

Considérez l'infrastructure comme un code.

politique basée sur l'identité

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

application inactive

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

IIo T

Voir Internet industriel des objets.

infrastructure immuable

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

VPC entrant (d'entrée)

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

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

migration incrémentielle

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

Industry 4.0

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

infrastructure

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

infrastructure en tant que code (IaC)

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

Internet industriel des objets (IIoT)

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

VPC d'inspection

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

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

Internet des objets (IoT)

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

interprétabilité

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

IoT

Voir Internet des objets.

Bibliothèque d'informations informatiques (ITIL)

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

gestion des services informatiques (ITSM)

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

ITIL

Consultez la bibliothèque d'informations informatiques.

ITSM

Voir Gestion des services informatiques.

L

contrôle d'accès basé sur des étiquettes (LBAC)

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

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

zone de destination

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

grand modèle de langage (LLM)

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

migration de grande envergure

Migration de 300 serveurs ou plus.

LBAC

Voir contrôle d'accès basé sur des étiquettes.

principe de moindre privilège

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

lift and shift

Voir 7 Rs.

système de poids faible

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

LLM

Voir le grand modèle de langage.

environnements inférieurs

Voir environnement.

M

machine learning (ML)

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

branche principale

Voir succursale.

malware

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

services gérés

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

système d'exécution de la fabrication (MES)

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

MAP

Voir Migration Acceleration Program.

mécanisme

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

compte membre

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

MAILLES

Voir le système d'exécution de la fabrication.

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

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

microservice

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

architecture de microservices

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

Programme d'accélération des migrations (MAP)

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

migration à grande échelle

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

usine de migration

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

métadonnées de migration

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

modèle de migration

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

Évaluation du portefeuille de migration (MPA)

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

Évaluation de la préparation à la migration (MRA)

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

stratégie de migration

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

ML

Voir apprentissage automatique.

modernisation

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

évaluation de la préparation à la modernisation

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

applications monolithiques (monolithes)

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

MPA

Voir Évaluation du portefeuille de migration.

MQTT

Voir Message Queuing Telemetry Transport.

classification multi-classes

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

infrastructure mutable

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

O

OAC

Voir Contrôle d'accès à l'origine.

OAI

Voir l'identité d'accès à l'origine.

OCM

Voir gestion du changement organisationnel.

migration hors ligne

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

OI

Consultez la section Intégration des opérations.

OLA

Voir l'accord au niveau opérationnel.

migration en ligne

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

OPC-UA

Voir Open Process Communications - Architecture unifiée.

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

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

accord au niveau opérationnel (OLA)

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

examen de l'état de préparation opérationnelle (ORR)

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

technologie opérationnelle (OT)

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

intégration des opérations (OI)

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

journal de suivi d'organisation

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

gestion du changement organisationnel (OCM)

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

contrôle d'accès d'origine (OAC)

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

identité d'accès d'origine (OAI)

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

ORR

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

DE

Voir technologie opérationnelle.

VPC sortant (de sortie)

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

P

limite des autorisations

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

informations personnelles identifiables (PII)

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

PII

Voir les informations personnelles identifiables.

manuel stratégique

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

PLC

Voir contrôleur logique programmable.

PLM

Consultez la section Gestion du cycle de vie des produits.

policy

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

persistance polyglotte

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

évaluation du portefeuille

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

predicate

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

prédicat pushdown

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

contrôle préventif

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

principal

Entité AWS capable d'effectuer des actions et d'accéder aux ressources. Cette entité est généralement un utilisateur root pour un Compte AWS rôle IAM ou un utilisateur. Pour plus

d'informations, veuillez consulter la rubrique Principal dans Termes et concepts relatifs aux rôles, dans la documentation IAM.

confidentialité dès la conception

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

zones hébergées privées

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

contrôle proactif

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

gestion du cycle de vie des produits (PLM)

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

environnement de production

Voir environnement.

contrôleur logique programmable (PLC)

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

chaînage rapide

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

pseudonymisation

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

publish/subscribe (pub/sub)

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

Q

plan de requête

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

régression du plan de requêtes

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

R

Matrice RACI

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

RAG

Voir Retrieval Augmented Generation.

rançongiciel

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

Matrice RASCI

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

RCAC

Voir contrôle d'accès aux lignes et aux colonnes.

réplica en lecture

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

réarchitecte

Voir 7 Rs.

objectif de point de récupération (RPO)

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

objectif de temps de récupération (RTO)

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

refactoriser

Voir 7 Rs.

Région

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

régression

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

réhéberger

Voir 7 Rs.

version

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

déplacer

Voir 7 Rs.

replateforme

Voir 7 Rs.

rachat

Voir 7 Rs.

résilience

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

politique basée sur les ressources

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

matrice responsable, redevable, consulté et informé (RACI)

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

contrôle réactif

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

retain

Voir 7 Rs.

se retirer

Voir 7 Rs.

Génération augmentée de récupération (RAG)

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

rotation

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

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

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

RPO

Voir l'objectif du point de récupération.

RTO

Voir l'objectif en matière de temps de rétablissement.

runbook

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

S

SAML 2.0

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

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

SCADA

Voir Contrôle de supervision et acquisition de données.

SCP

Voir la politique de contrôle des services.

secret

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

sécurité dès la conception

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

contrôle de sécurité

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

renforcement de la sécurité

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

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

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

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

automatisation des réponses de sécurité

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

chiffrement côté serveur

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

Politique de contrôle des services (SCP)

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

point de terminaison du service

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

contrat de niveau de service (SLA)

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

indicateur de niveau de service (SLI)

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

objectif de niveau de service (SLO)

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

modèle de responsabilité partagée

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

SIEM

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

point de défaillance unique (SPOF)

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

SLA

Voir le contrat de niveau de service.

SLI

Voir l'indicateur de niveau de service.

SLO

Voir l'objectif de niveau de service.

split-and-seed modèle

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

SPOF

Voir point de défaillance unique.

schéma en étoile

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

modèle de figuier étrangleur

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

sous-réseau

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

contrôle de supervision et acquisition de données (SCADA)

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

chiffrement symétrique

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

tests synthétiques

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

invite du système

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

T

tags

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

variable cible

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

liste de tâches

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

environnement de test

Voir environnement.

entraînement

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

passerelle de transit

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

flux de travail basé sur jonction

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

accès sécurisé

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

section Utilisation AWS Organizations avec d'autres AWS services dans la AWS Organizations documentation.

réglage

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

équipe de deux pizzas

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

U

incertitude

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

tâches indifférenciées

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

environnements supérieurs

Voir environnement.

V

mise à vide

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

contrôle de version

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

Appairage de VPC

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

vulnérabilités

Défaut logiciel ou matériel qui compromet la sécurité du système.

W

cache actif

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

données chaudes

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

fonction de fenêtre

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

charge de travail

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

flux de travail

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

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

VER

Voir écrire une fois, lire plusieurs.

WQF

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

écrire une fois, lire plusieurs (WORM)

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

Z

exploit Zero-Day

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

vulnérabilité « jour zéro »

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

invite Zero-Shot

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

application zombie

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

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