Récupérez les options et architectures de génération augmentée sur AWS

AWS Directives prescriptives

AWS Directives prescriptives: Récupérez les options et architectures de génération augmentée sur AWS

Table of Contents

Récupérez les options et architectures de génération augmentée sur AWS

Mithil Shah, Rajeev Muralidhar et Natacha Fort, Amazon Web Services

octobre 2024 (historique du document)

L'IA générative fait référence à un sous-ensemble de modèles d'IA capables de créer de nouveaux contenus et artefacts, tels que des images, des vidéos, du texte et du son, à partir d'une simple invite textuelle. Les modèles d'IA générative sont entraînés sur de grandes quantités de données qui englobent un large éventail de sujets et de tâches. Cela leur permet de faire preuve d'une polyvalence remarquable dans l'exécution de diverses tâches, même celles pour lesquelles ils n'ont pas reçu de formation explicite. En raison de la capacité d'un seul modèle à effectuer plusieurs tâches, ces modèles sont souvent appelés modèles de base (FMs).

L'une des applications remarquables des modèles d'IA générative est leur capacité à répondre aux questions. Cependant, des défis spécifiques se posent lorsque ces modèles sont utilisés pour répondre à des questions basées sur des documents personnalisés. Les documents personnalisés peuvent inclure des informations exclusives, des sites Web internes, de la documentation interne, des SharePoint pages, des pages et autres. Confluence L'une des options consiste à utiliser la génération augmentée de récupération (RAG). Avec RAG, le modèle de base référence une source de données faisant autorité qui se trouve en dehors de ses sources de données de formation (telles que vos documents personnalisés) avant de générer une réponse.

Ce guide décrit les différentes options d'IA générative disponibles pour répondre aux questions de la documentation personnalisée, notamment les systèmes RAG (Retrieval Augmented Generation). Il fournit également une vue d'ensemble de la création de systèmes RAG sur Amazon Web Services (AWS). En passant en revue les options et architectures RAG, vous pouvez choisir entre des services entièrement gérés sur des architectures RAG AWS et des architectures RAG personnalisées.

Public visé

Ce guide s'adresse aux architectes et aux gestionnaires d'IA générative qui souhaitent créer une solution RAG, examiner les architectures disponibles et comprendre les avantages et les inconvénients de chaque option.

Objectifs

Ce guide vous aide à accomplir les tâches suivantes :

- Découvrez les options d'IA générative disponibles pour répondre aux questions à partir de documents personnalisés
- Passez en revue les options d'architecture pour les systèmes RAG sur AWS
- Comprenez les avantages et les inconvénients de chaque option RAG
- Choisissez une architecture RAG adaptée à votre environnement AWS

Options d'IA génératives pour interroger des documents personnalisés

Organisations disposent souvent de différentes sources de données structurées et non structurées. Ce guide explique comment utiliser l'IA générative pour répondre à des questions à partir de données non structurées.

Les données non structurées de votre organisation peuvent provenir de différentes sources. Il peut s'agir de fichiers texte PDFs, de wikis internes, de documents techniques, de sites Web destinés au public, de bases de connaissances ou autres. Si vous souhaitez un modèle de base capable de répondre aux questions relatives aux données non structurées, les options suivantes sont disponibles :

- Formez un nouveau modèle de base en utilisant vos documents personnalisés et d'autres données de formation
- Affinez un modèle de base existant en utilisant les données de vos documents personnalisés
- Utilisez l'apprentissage contextuel pour transmettre un document au modèle de base lorsque vous posez une question
- Utiliser une approche RAG (Retrieval Augmented Generation)

Former un nouveau modèle de base à partir de zéro qui inclut vos données personnalisées est une entreprise ambitieuse. Quelques entreprises l'ont fait avec succès, notamment Bloomberg avec leur BloombergGPTmodèle. Un autre exemple est le EXAONEmodèle multimodal deLG AI Research, qui a été formé en utilisant 600 milliards d'œuvres d'art et 250 millions d'images haute résolution, accompagnées de texte. Selon The Cost of AI : Should You Build or Buy Your Foundation Model (LinkedIn), un modèle similaire Meta Llama 2 coûte environ 4,8 millions de dollars américains à former. Deux conditions principales sont requises pour former un modèle à partir de zéro : l'accès aux ressources (financières, techniques, temps) et un retour sur investissement clair. Si cela ne vous convient pas, l'option suivante consiste à peaufiner un modèle de fondation existant.

Pour affiner un modèle existant, il faut prendre un modèle, tel qu'un modèle Amazon Titan, Mistral ou Llama, puis l'adapter à vos données personnalisées. Il existe différentes techniques de réglage précis, dont la plupart impliquent de ne modifier que quelques paramètres au lieu de modifier tous les paramètres du modèle. C'est ce qu'on appelle un réglage fin efficace en fonction des paramètres. Il existe deux méthodes principales pour affiner les réglages :

- Le réglage fin supervisé utilise des données étiquetées et vous aide à entraîner le modèle pour un nouveau type de tâche. Par exemple, si vous souhaitez générer un rapport basé sur un formulaire PDF, vous devrez peut-être apprendre au modèle à le faire en fournissant suffisamment d'exemples.
- Le réglage fin non supervisé est indépendant des tâches et adapte le modèle de base à vos propres données. Il entraîne le modèle pour qu'il comprenne le contexte de vos documents. Le modèle affiné crée ensuite du contenu, tel qu'un rapport, en utilisant un style plus personnalisé pour votre organisation.

Cependant, le réglage précis n'est peut-être pas idéal pour les cas d'utilisation sous forme de questions-réponses. Pour plus d'informations, consultez la section Comparaison entre RAG et optimisation dans ce guide.

Lorsque vous posez une question, vous pouvez transmettre à un document le modèle de base et utiliser l'apprentissage contextuel du modèle pour renvoyer les réponses du document. Cette option convient à l'interrogation ad hoc d'un seul document. Cependant, cette solution ne fonctionne pas bien pour interroger plusieurs documents ou pour interroger des systèmes et des applications, tels que Microsoft SharePoint ou Atlassian Confluence.

La dernière option consiste à utiliser RAG. Avec RAG, le modèle de base référence vos documents personnalisés avant de générer une réponse. RAG étend les capacités du modèle à la base de connaissances interne de votre organisation, le tout sans qu'il soit nécessaire de modifier le modèle. Il s'agit d'une approche rentable pour améliorer le résultat du modèle afin qu'il reste pertinent, précis et utile dans divers contextes.

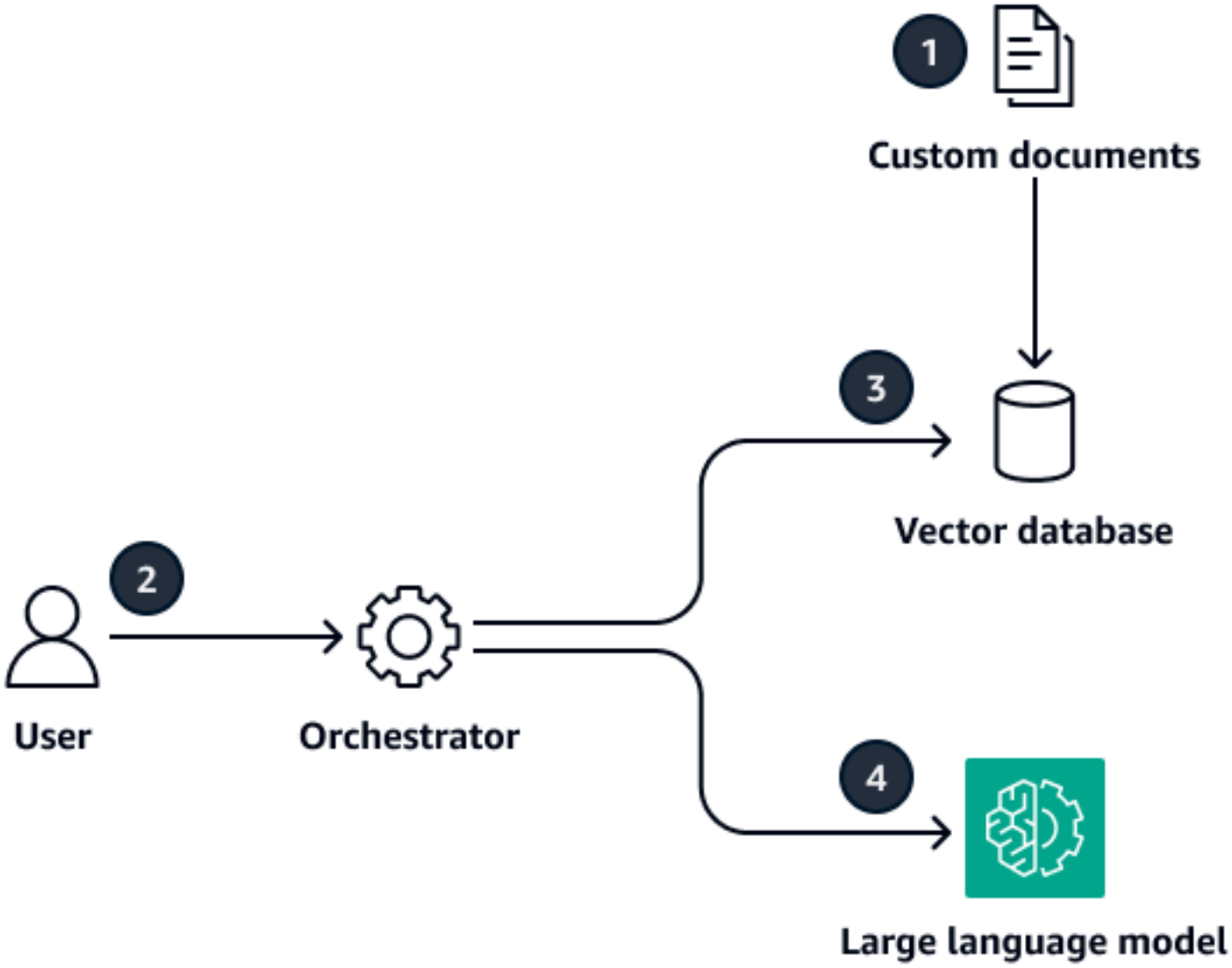
Rubriques de cette section :

- Comprendre la génération augmentée de récupération
- Comparaison entre la récupération, la génération augmentée et le réglage précis
- Cas d'utilisation de Retrieval Augmented Generation

Comprendre la génération augmentée de récupération

La génération augmentée de récupération (RAG) est une technique utilisée pour compléter un grand modèle de langage (LLM) avec des données externes, telles que les documents internes d'une entreprise. Cela fournit au modèle le contexte dont il a besoin pour produire des résultats précis et utiles pour votre cas d'utilisation spécifique. RAG est une approche pragmatique et efficace à utiliser

LLMs dans une entreprise. Le schéma suivant donne un aperçu général du fonctionnement d'une approche RAG.



D'une manière générale, le processus RAG comporte quatre étapes. La première étape est effectuée une fois, et les trois autres étapes sont effectuées autant de fois que nécessaire :

1. Vous créez des intégrations pour intégrer les documents internes dans une base de données vectorielle. Les intégrations sont des représentations numériques du texte dans les documents qui capturent la signification sémantique ou contextuelle des données. Une base de données vectorielle est essentiellement une base de données de ces intégrations, parfois appelée magasin de vecteurs ou index vectoriel. Cette étape nécessite le nettoyage, le formatage et le découpage des données, mais il s'agit d'une activité initiale ponctuelle.

2. Un humain soumet une requête en langage naturel.
3. Un orchestrateur effectue une recherche de similarité dans la base de données vectorielle et récupère les données pertinentes. L'orchestrateur ajoute les données récupérées (également appelées contexte) à l'invite contenant la requête.
4. L'orchestrateur envoie la requête et le contexte au LLM. Le LLM génère une réponse à la requête en utilisant le contexte supplémentaire.

Du point de vue de l'utilisateur, RAG ressemble à une interaction avec n'importe quel LLM. Cependant, le système en sait beaucoup plus sur le contenu en question et fournit des réponses adaptées à la base de connaissances de l'organisation.

Pour plus d'informations sur le fonctionnement d'une approche RAG, voir Qu'est-ce que RAG sur le AWS site Web.

Composants des systèmes RAG destinés à la production

La création d'un système RAG au niveau de la production nécessite de réfléchir à plusieurs aspects différents du flux de travail RAG. Conceptuellement, un flux de travail RAG au niveau de la production nécessite les fonctionnalités et composants suivants, quelle que soit l'implémentation spécifique :

- Connecteurs : ils connectent différentes sources de données d'entreprise à la base de données vectorielle. Les bases de données transactionnelles et analytiques sont des exemples de sources de données structurées. Les exemples de sources de données non structurées incluent les magasins d'objets, les bases de code et les plateformes SaaS (Software as a Service). Chaque source de données peut nécessiter des modèles de connectivité, des licences et des configurations différents.
- Traitement des données — Les données se présentent sous de nombreuses formes, telles que des images numérisées PDFs, des documents, des présentations et des Microsoft SharePoint fichiers. Vous devez utiliser des techniques de traitement des données pour extraire, traiter et préparer les données pour l'indexation.
- Intégrations — Pour effectuer une recherche de pertinence, vous devez convertir vos documents et les requêtes des utilisateurs dans un format compatible. En utilisant des modèles linguistiques intégrés, vous convertissez les documents en représentation numérique. Il s'agit essentiellement d'entrées pour le modèle de base sous-jacent.
- Base de données vectorielle — La base de données vectorielle est un index des intégrations, du texte associé et des métadonnées. L'index est optimisé pour la recherche et l'extraction.

- Retriever : pour la requête de l'utilisateur, le récupérateur extrait le contexte pertinent dans la base de données vectorielle et classe les réponses en fonction des besoins de l'entreprise.
- Modèle de base — Le modèle de base d'un système RAG est généralement un LLM. En traitant le contexte et l'invite, le modèle de base génère et met en forme une réponse pour l'utilisateur.
- Garde-corps — Les garde-corps sont conçus pour garantir que la requête, le contexte rapide, récupéré et la réponse LLM sont précis, responsables, éthiques et exempts d'hallucinations et de préjugés.
- Orchestrateur — L'orchestrateur est responsable de la planification et de la gestion du end-to-end flux de travail.
- Expérience utilisateur — Généralement, l'utilisateur interagit avec une interface de chat conversationnelle dotée de fonctionnalités riches, notamment l'affichage de l'historique des discussions et la collecte des commentaires des utilisateurs sur les réponses.
- Gestion des identités et des utilisateurs — Il est essentiel de contrôler l'accès des utilisateurs à l'application de manière précise. Dans le AWS Cloud, les politiques, les rôles et les autorisations sont généralement gérés via Gestion des identités et des accès AWS (IAM).

De toute évidence, la planification, le développement, la publication et la gestion d'un système RAG nécessitent beaucoup de travail. Des services entièrement gérés, tels qu'Amazon Bedrock ou Amazon Q Business, peuvent vous aider à gérer une partie des tâches lourdes indifférenciées. Cependant, les architectures RAG personnalisées peuvent fournir un meilleur contrôle sur les composants, tels que le récupérateur ou la base de données vectorielle.

Comparaison entre la récupération, la génération augmentée et le réglage précis

Le tableau suivant décrit les avantages et les inconvénients des approches de réglage fin et basées sur le RAG.

Approche	Avantages	Inconvénients
Peaufinage	• Si un modèle affiné est formé à l'aide d'une approche non supervisé e, il est alors en mesure	• La mise au point peut prendre de quelques heures à quelques jours, selon la taille du modèle. Ce

Approche	Avantages	Inconvénients
	de créer du contenu qui correspond mieux au style de votre organisation. • Un modèle affiné basé sur des données propriétaires ou réglementaires peut aider votre entreprise à respecter les normes internes ou sectorielles en matière de données et de conformité.	n'est donc pas une bonne solution si vos documents personnalisés changent fréquemment. • Le réglage précis nécessite la compréhension de techniques telles que l'adaptation de bas rang (LoRa) et le réglage précis efficace des paramètres (PEFT). Le réglage précis peut nécessiter l'intervention d'un data scientist. • Il se peut que le réglage précis ne soit pas disponible pour tous les modèles. • Les modèles affinés ne fournissent aucune référence à la source dans leurs réponses. • Il peut y avoir un risque accru d'hallucination lorsque l'on utilise un modèle affiné pour répondre à des questions.

Approche	Avantages	Inconvénients
RAG	• RAG vous permet de créer un système de réponse aux questions pour vos documents personnalisés sans avoir à le peaufiner. • RAG peut intégrer les derniers documents en quelques minutes. • AWS propose des solutions RAG entièrement gérées. Par conséquent, aucun data scientist ni aucune connaissance spécialisée en apprentissage automatiq ue ne sont nécessaires. • Dans sa réponse, un modèle RAG fournit une référence à la source d'information. • Comme RAG utilise le contexte de la recherche vectorielle comme base de la réponse générée, le risque d'hallucination est réduit.	• RAG ne fonctionne pas bien lorsqu'il s'agit de résumer des informations provenant de documents entiers.

Si vous devez créer une solution de réponse aux questions qui fait référence à vos documents personnalisés, nous vous recommandons de commencer par une approche basée sur le RAG. Utilisez le réglage fin si vous avez besoin que le modèle exécute des tâches supplémentaires, telles que la synthèse.

Vous pouvez combiner les approches de réglage fin et RAG dans un seul modèle. Dans ce cas, l'architecture RAG ne change pas, mais le LLM qui génère la réponse est également affiné avec les

documents personnalisés. Cela combine le meilleur des deux mondes et constitue peut-être une solution optimale pour votre cas d'utilisation. Pour plus d'informations sur la manière de combiner le réglage fin supervisé avec le RAG, consultez la recherche RAFT : Adapting Language Model to Domain Specific RAG du. University of California, Berkeley

Cas d'utilisation de Retrieval Augmented Generation

Les cas d'utilisation courants d'une approche RAG sont les suivants :

- Moteurs de recherche — Les moteurs de recherche utilisant le tag peuvent fournir des extraits plus précis et plus up-to-date détaillés dans leurs résultats de recherche.
- Systèmes de réponse aux questions — RAG peut améliorer la qualité des réponses dans les systèmes de réponse aux questions. Le modèle basé sur la récupération utilise la recherche par similarité pour trouver des passages ou des documents pertinents contenant la réponse. Il génère ensuite une réponse concise et pertinente sur la base de ces informations.
- Commerce de détail ou commerce électronique — RAG peut améliorer l'expérience utilisateur dans le commerce électronique en fournissant des recommandations de produits plus pertinentes et personnalisées. En récupérant et en incorporant des informations sur les préférences des utilisateurs et les détails des produits, RAG peut générer des recommandations plus précises et plus utiles pour les clients.
- Industriel ou manufacturier — Dans le domaine de la fabrication, RAG vous aide à accéder rapidement aux informations critiques, telles que les opérations des usines. Il peut également contribuer aux processus de prise de décision, au dépannage et à l'innovation organisationnelle. Pour les fabricants qui opèrent dans des cadres réglementaires stricts, RAG peut rapidement consulter les réglementations et les normes de conformité mises à jour auprès de sources internes et externes, telles que les normes industrielles ou les agences de réglementation.
- Santé — RAG a du potentiel dans le secteur de la santé, où l'accès à des informations précises et opportunes est crucial. En récupérant et en incorporant les connaissances médicales pertinentes provenant de sources externes, RAG peut fournir des réponses plus précises et contextuelles dans les applications de santé. De telles applications augmentent les informations accessibles par un clinicien humain, qui prend en fin de compte l'appel et non le modèle.
- Juridique — Le RAG peut être appliqué de manière efficace dans des scénarios juridiques, tels que les fusions et acquisitions, où des documents juridiques complexes fournissent un contexte pour les requêtes. Cela peut aider les professionnels du droit à résoudre rapidement des problèmes réglementaires complexes.

Options de génération augmentée de récupération entièrement gérées sur AWS

Pour gérer les flux de travail RAG (Retrieval Augmented Generation) sur AWS, vous pouvez utiliser des pipelines RAG personnalisés ou utiliser certaines des fonctionnalités de services entièrement gérés proposées. AWS Parce qu'ils incluent de nombreux composants de base d'un système basé sur RAG, les services entièrement gérés peuvent vous aider à gérer certaines tâches lourdes indifférenciées. Cependant, ces services offrent moins de possibilités de personnalisation.

Les solutions entièrement gérées Services AWS utilisent des connecteurs pour ingérer des données provenant de sources de données externes, telles que des sites Web, Atlassian Confluence ou Microsoft. SharePoint Les sources de données prises en charge varient selon Service AWS.

Cette section explore les options entièrement gérées suivantes pour créer des flux de travail RAG sur AWS :

- Bases de connaissances pour Amazon Bedrock
- Amazon Q Business
- Amazon SageMaker AI Canvas

Pour plus d'informations sur le choix entre ces options, consultez Choix d'une option de génération augmentée de récupération sur AWS ce guide.

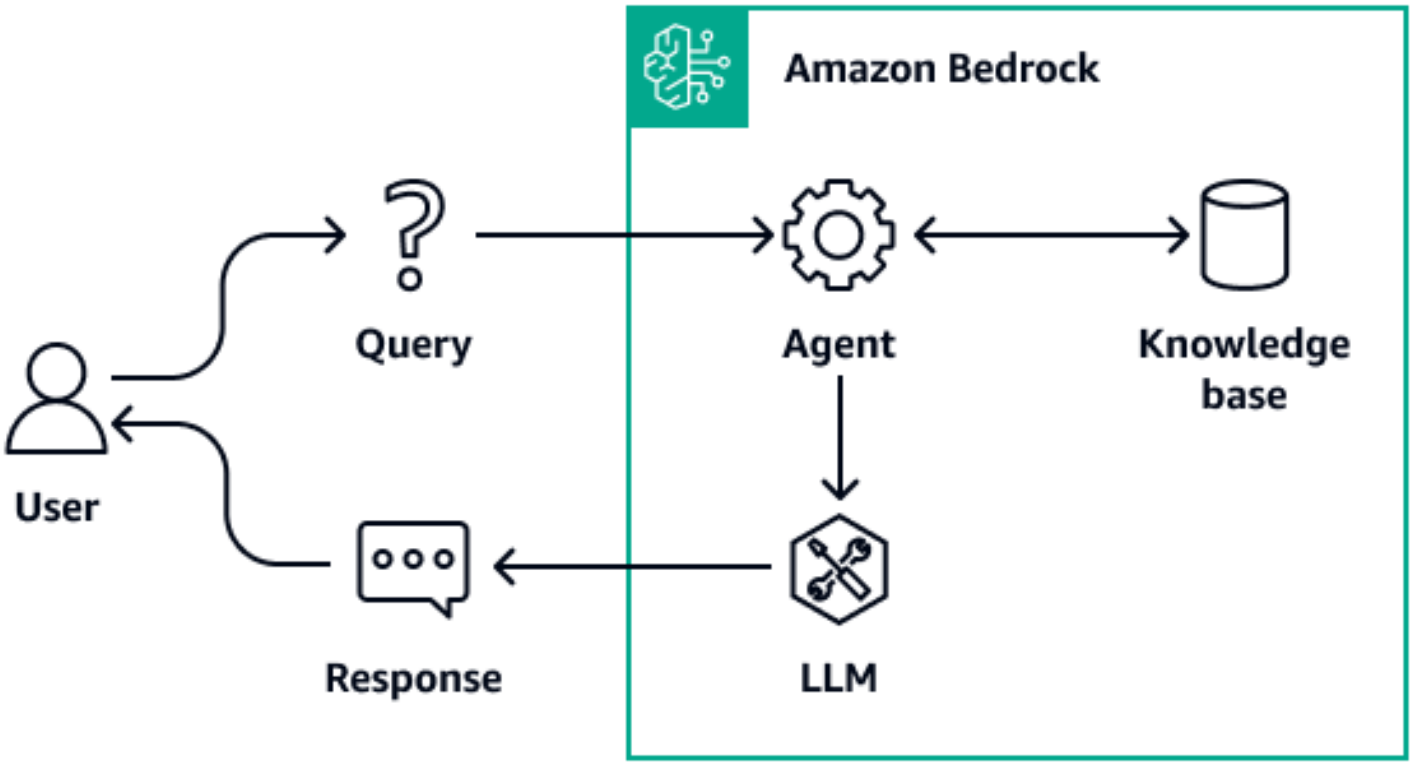
Bases de connaissances pour Amazon Bedrock

Amazon Bedrock est un service entièrement géré qui met à votre disposition des modèles de base très performants (FMs) issus des principales startups d'IA et d'Amazon via une API unifiée. Les bases de connaissances sont une fonctionnalité d'Amazon Bedrock qui vous permet de mettre en œuvre l'intégralité du flux de travail RAG, de l'ingestion à la récupération et à l'augmentation rapide. Il n'est pas nécessaire de créer des intégrations personnalisées aux sources de données ou de gérer les flux de données. La gestion du contexte de session est intégrée afin que votre application d'IA générative puisse facilement prendre en charge des conversations à plusieurs tours.

Une fois que vous avez indiqué l'emplacement de vos données, les bases de connaissances d'Amazon Bedrock récupèrent les documents en interne, les fragmentent en blocs de texte, les convertissent en éléments incorporés, puis stockent les éléments intégrés dans la base de données

vectorielle de votre choix. Amazon Bedrock gère et met à jour les intégrations, en synchronisant la base de données vectorielle avec les données. Pour plus d'informations sur le fonctionnement des bases de connaissances, consultez Comment fonctionnent les bases de connaissances Amazon Bedrock.

Si vous ajoutez des bases de connaissances à un agent Amazon Bedrock, celui-ci identifie la base de connaissances appropriée en fonction des informations saisies par l'utilisateur. L'agent récupère les informations pertinentes et les ajoute à l'invite de saisie. L'invite mise à jour fournit au modèle davantage d'informations contextuelles pour générer une réponse. Pour améliorer la transparence et minimiser les hallucinations, les informations extraites de la base de connaissances sont traçables jusqu'à leur source.



Amazon Bedrock prend en charge les deux solutions suivantes APIs pour RAG :

- RetrieveAndGenerate— Vous pouvez utiliser cette API pour interroger votre base de connaissances et générer des réponses à partir des informations qu'elle récupère. En interne, Amazon Bedrock convertit les requêtes en intégrations, interroge la base de connaissances, ajoute les résultats de recherche sous forme d'informations contextuelles à l'invite et renvoie la réponse générée par le LLM. Amazon Bedrock gère également la mémoire à court terme de la conversation afin de fournir des résultats plus contextuels.

- Récupérer : vous pouvez utiliser cette API pour interroger votre base de connaissances à l'aide d'informations extraites directement de la base de connaissances. Vous pouvez utiliser les informations renvoyées par cette API pour traiter le texte extrait, évaluer sa pertinence ou développer un flux de travail distinct pour la génération de réponses. En interne, Amazon Bedrock convertit les requêtes en intégrations, effectue des recherches dans la base de connaissances et renvoie les résultats pertinents. Vous pouvez créer des flux de travail supplémentaires en plus des résultats de recherche. Par exemple, vous pouvez utiliser le LangChain`AmazonKnowledgeBasesRetriever`plugin pour intégrer les flux de travail RAG dans des applications d'IA générative.

Pour obtenir des exemples de modèles architecturaux et des step-by-step instructions d'utilisation APIs, consultez Knowledge Bases qui propose désormais une expérience RAG entièrement gérée dans Amazon Bedrock (article de AWS blog). Pour plus d'informations sur l'utilisation de l'`RetrieveAndGenerate`API afin de créer un flux de travail RAG pour une application intelligente basée sur le chat, consultez Création d'une application de chatbot contextuelle à l'aide des bases de connaissances Amazon Bedrock (article de blog).AWS

Sources de données pour les bases de connaissances

Vous pouvez connecter vos données propriétaires à une base de connaissances. Après avoir configuré un connecteur de source de données, vous pouvez synchroniser ou maintenir vos données à jour avec votre base de connaissances et les rendre disponibles pour les requêtes. Les bases de connaissances Amazon Bedrock prennent en charge les connexions aux sources de données suivantes :

- Amazon Simple Storage Service (Amazon S3) — Vous pouvez connecter un bucket Amazon S3 à une base de connaissances Amazon Bedrock à l'aide de la console ou de l'API. La base de connaissances ingère et indexe les fichiers du bucket. Ce type de source de données prend en charge les fonctionnalités suivantes :
  - Champs de métadonnées du document : vous pouvez inclure un fichier distinct pour spécifier les métadonnées des fichiers du compartiment Amazon S3. Vous pouvez ensuite utiliser ces champs de métadonnées pour filtrer et améliorer la pertinence des réponses.
  - Filtres d'inclusion ou d'exclusion : vous pouvez inclure ou exclure certains contenus lors de l'exploration.
  - Synchronisation incrémentielle : les modifications de contenu sont suivies et seul le contenu modifié depuis la dernière synchronisation est analysé.

- Confluence— Vous pouvez connecter une Atlassian Confluence instance à une base de connaissances Amazon Bedrock à l'aide de la console ou de l'API. Ce type de source de données prend en charge les fonctionnalités suivantes :
  - Détection automatique des principaux champs du document — Les champs de métadonnées sont automatiquement détectés et analysés. Vous pouvez utiliser ces champs pour le filtrage.
  - Filtres de contenu d'inclusion ou d'exclusion : vous pouvez inclure ou exclure certains contenus en utilisant un préfixe ou un modèle d'expression régulière sur l'espace, le titre de la page, le titre du blog, le commentaire, le nom de la pièce jointe ou l'extension.
  - Synchronisation incrémentielle : les modifications de contenu sont suivies et seul le contenu modifié depuis la dernière synchronisation est analysé.
  - OAuth Authentification 2.0, authentification avec jeton Confluence API — Les informations d'authentification sont stockées dans AWS Secrets Manager.
- Microsoft SharePoint— Vous pouvez connecter une SharePoint instance à une base de connaissances à l'aide de la console ou de l'API. Ce type de source de données prend en charge les fonctionnalités suivantes :
  - Détection automatique des principaux champs du document — Les champs de métadonnées sont automatiquement détectés et analysés. Vous pouvez utiliser ces champs pour le filtrage.
  - Filtres de contenu d'inclusion ou d'exclusion : vous pouvez inclure ou exclure certains contenus en utilisant un préfixe ou un modèle d'expression régulière sur le titre de la page principale, le nom de l'événement et le nom du fichier (y compris son extension).
  - Synchronisation incrémentielle : les modifications de contenu sont suivies et seul le contenu modifié depuis la dernière synchronisation est analysé.
  - OAuth Authentification 2.0 — Les informations d'authentification sont stockées dans AWS Secrets Manager.
- Salesforce— Vous pouvez connecter une Salesforce instance à une base de connaissances à l'aide de la console ou de l'API. Ce type de source de données prend en charge les fonctionnalités suivantes :
  - Détection automatique des principaux champs du document — Les champs de métadonnées sont automatiquement détectés et analysés. Vous pouvez utiliser ces champs pour le filtrage.
  - Filtres de contenu d'inclusion ou d'exclusion : vous pouvez inclure ou exclure certains contenus à l'aide d'un préfixe ou d'un modèle d'expression régulière. Pour obtenir la liste des types de contenu auxquels vous pouvez appliquer des filtres, consultez les filtres d'inclusion/exclusion dans la documentation Amazon Bedrock.

  - Synchronisation incrémentielle : les modifications de contenu sont suivies et seul le contenu modifié depuis la dernière synchronisation est analysé.
  - OAuth Authentification 2.0 — Les informations d'authentification sont stockées dans AWS Secrets Manager.
- Web Crawler — Un robot d'exploration Web Amazon Bedrock se connecte aux données que vous fournissez et les URLs explore. Les fonctionnalités suivantes sont prises en charge :
  - Sélectionnez plusieurs URLs à explorer
  - Respectez les directives standard du fichier robots.txt, telles que `Allow` et `Disallow`
  - Exclure URLs ceux qui correspondent à un modèle
  - Limitez le taux de rampage
  - Dans Amazon CloudWatch, consultez le statut de chaque URL analysée

Pour plus d'informations sur les sources de données que vous pouvez connecter à votre base de connaissances Amazon Bedrock, consultez Créer un connecteur de source de données pour votre base de connaissances.

Bases de données vectorielles pour les bases de connaissances

Lorsque vous établissez une connexion entre la base de connaissances et la source de données, vous devez configurer une base de données vectorielle, également appelée magasin de vecteurs. Une base de données vectorielle est l'endroit où Amazon Bedrock stocke, met à jour et gère les intégrations qui représentent vos données. Chaque source de données prend en charge différents types de bases de données vectorielles. Pour déterminer les bases de données vectorielles disponibles pour votre source de données, consultez les types de sources de données.

Si vous préférez qu'Amazon Bedrock crée automatiquement une base de données vectorielle dans Amazon OpenSearch Serverless pour vous, vous pouvez choisir cette option lors de la création de la base de connaissances. Toutefois, vous pouvez également choisir de configurer votre propre base de données vectorielles. Si vous configurez votre propre base de données vectorielles, reportez-vous à la section Conditions requises pour votre propre magasin de vecteurs pour obtenir une base de connaissances. Chaque type de base de données vectorielle possède ses propres prérequis.

En fonction de votre type de source de données, les bases de connaissances Amazon Bedrock prennent en charge les bases de données vectorielles suivantes :

- Amazon OpenSearch sans serveur

- Amazon Aurora PostgreSQL-Compatible Edition
- Pinecone(Pineconedocumentation)
- Redis Enterprise Cloud(Redisdocumentation)
- MongoDB Atlas(MongoDBdocumentation)

Amazon Q Business

Amazon Q Business est un assistant entièrement géré basé sur l'IA générative que vous pouvez configurer pour répondre aux questions, fournir des résumés, générer du contenu et effectuer des tâches en fonction des données de votre entreprise. Il permet aux utilisateurs finaux de recevoir des réponses immédiates, conformes aux autorisations, provenant de sources de données d'entreprise, accompagnées de citations.

Fonctions principales

Les fonctionnalités suivantes d'Amazon Q Business peuvent vous aider à créer une application d'IA générative basée sur RAG adaptée à la production :

- Connecteurs intégrés : Amazon Q Business prend en charge plus de 40 types de connecteurs, tels que les connecteurs pour Adobe Experience Manager (AEM)Salesforce,Jira, etMicrosoft SharePoint. Pour une liste complète, voir Connecteurs pris en charge. Si vous avez besoin d'un connecteur qui n'est pas pris en charge, vous pouvez utiliser Amazon AppFlow pour extraire les données de votre source de données vers Amazon Simple Storage Service (Amazon S3), puis connecter Amazon Q Business au compartiment Amazon S3. Pour obtenir la liste complète des sources de données prises AppFlow en charge par Amazon, consultez la section Applications prises en charge.
- Pipelines d'indexation intégrés — Amazon Q Business fournit un pipeline intégré pour indexer les données dans une base de données vectorielle. Vous pouvez utiliser une AWS Lambda fonction pour ajouter une logique de prétraitement à votre pipeline d'indexation.
- Options d'index : vous pouvez créer et approvisionner un index natif dans Amazon Q Business, et vous pouvez utiliser un récupérateur Amazon Q Business pour extraire les données de cet index. Vous pouvez également utiliser un index Amazon Kendra préconfiguré en tant que récupérateur. Pour plus d'informations, consultez Création d'un récupérateur pour une application Amazon Q Business.

- Modèles de base — Amazon Q Business utilise les modèles de base pris en charge dans Amazon Bedrock. Pour obtenir la liste complète, consultez la section Modèles de fondation pris en charge dans Amazon Bedrock.
- Plug-ins — Amazon Q Business permet d'utiliser des plug-ins pour s'intégrer aux systèmes cibles, par exemple une méthode automatisée de synthèse des informations sur les tickets et de création de tickets dansJira. Une fois configurés, les plug-ins peuvent prendre en charge des actions de lecture et d'écriture qui peuvent vous aider à améliorer la productivité des utilisateurs finaux. Amazon Q Business prend en charge deux types de plug-ins : les plug-ins intégrés et les plug-ins personnalisés.
- Garde-corps — Amazon Q Business prend en charge les contrôles globaux et les contrôles thématiques. Par exemple, ces contrôles peuvent détecter des informations personnelles identifiables (PII), des abus ou des informations sensibles dans les invites. Pour plus d'informations, consultez la section Contrôles administratifs et barrières de sécurité dans Amazon Q Business.
- Gestion des identités — Avec Amazon Q Business, vous pouvez gérer les utilisateurs et leur accès à l'application d'IA générative basée sur RAG. Pour plus d'informations, consultez la section Gestion des identités et des accès pour Amazon Q Business. En outre, les connecteurs Amazon Q Business indexent les informations de la liste de contrôle d'accès (ACL) jointes à un document en même temps que le document lui-même. Amazon Q Business stocke ensuite les informations ACL qu'il indexe dans l'Amazon Q Business User Store afin de créer des mappages d'utilisateurs et de groupes et de filtrer les réponses au chat en fonction de l'accès de l'utilisateur final aux documents. Pour plus d'informations, consultez la section Concepts du connecteur de source de données.
- Enrichissement des documents : la fonctionnalité d'enrichissement des documents vous permet de contrôler à la fois les documents et les attributs de document qui sont ingérés dans votre index, ainsi que la manière dont ils sont ingérés. Pour ce faire, deux approches peuvent être utilisées :
  - Configurer les opérations de base : utilisez les opérations de base pour ajouter, mettre à jour ou supprimer des attributs de document dans vos données. Par exemple, vous pouvez effacer les données personnelles en choisissant de supprimer tous les attributs de document liés aux informations personnelles.
  - Configuration des fonctions Lambda : utilisez une fonction Lambda préconfigurée pour appliquer une logique de manipulation des attributs de document plus personnalisée et avancée à vos données. Par exemple, les données de votre entreprise peuvent être stockées sous forme d’images numérisées. Dans ce cas, vous pouvez utiliser une fonction Lambda pour exécuter la reconnaissance optique de caractères (OCR) sur les documents numérisés afin d'en extraire du texte. Chaque document numérisé est ensuite traité comme un document texte lors de

l'ingestion. Enfin, pendant le chat, Amazon Q prend en compte les données textuelles extraites des documents numérisés lorsqu'il génère des réponses.

Lorsque vous implémentez votre solution, vous pouvez choisir de combiner les deux approches d'enrichissement des documents. Vous pouvez utiliser des opérations de base pour effectuer une première analyse de vos données, puis utiliser une fonction Lambda pour des opérations plus complexes. Pour plus d'informations, consultez la section Enrichissement des documents dans Amazon Q Business.

- Intégration — Après avoir créé votre application Amazon Q Business, vous pouvez l'intégrer à d'autres applications, telles que Slack ouMicrosoft Teams. Par exemple, consultez Déployer une Slack passerelle pour Amazon Q Business et Déployer une Microsoft Teams passerelle pour Amazon Q Business (articles de AWS blog).

Personnalisation pour l'utilisateur final

Amazon Q Business prend en charge le téléchargement de documents susceptibles de ne pas être stockés dans les sources de données et dans l'index de votre organisation. Les documents téléchargés ne sont pas stockés. Ils ne peuvent être utilisés que pour la conversation au cours de laquelle les documents sont téléchargés. Amazon Q Business prend en charge le téléchargement de types de documents spécifiques. Pour plus d'informations, consultez Importer des fichiers et discuter dans Amazon Q Business.

Amazon Q Business inclut une fonctionnalité de filtrage par attribut de document. Les administrateurs et les utilisateurs finaux peuvent utiliser cette fonctionnalité. Les administrateurs peuvent personnaliser et contrôler les réponses au chat pour les utilisateurs finaux à l'aide d'attributs. Par exemple, si le type de source de données est un attribut attaché à vos documents, vous pouvez spécifier que les réponses au chat ne doivent être générées qu'à partir d'une source de données spécifique. Vous pouvez également autoriser les utilisateurs finaux à restreindre la portée des réponses au chat en utilisant les filtres d'attributs que vous avez sélectionnés.

Les utilisateurs finaux peuvent créer des applications Amazon Q légères et spécialement conçues au sein de votre environnement d'applications Amazon Q Business au sens large. Les applications Amazon Q permettent d'automatiser les tâches pour un domaine spécifique, par exemple une application spécialement conçue pour l'équipe marketing.

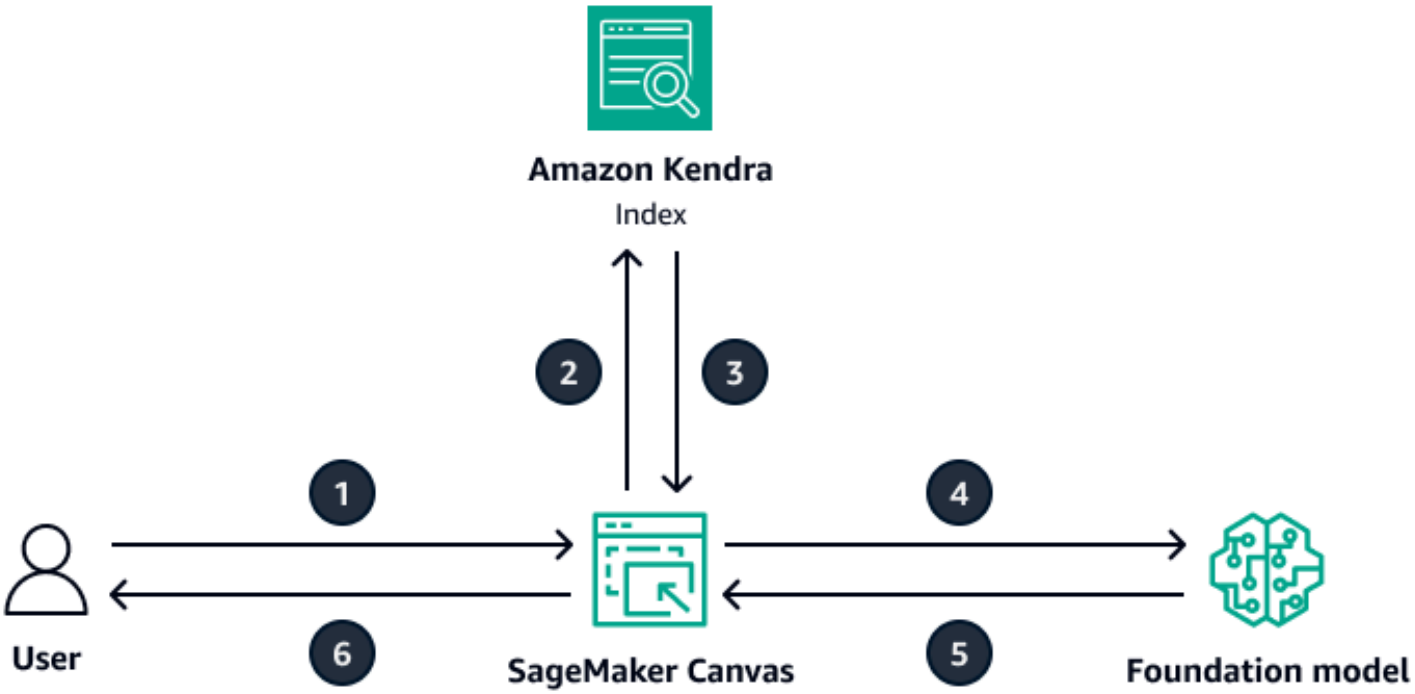
Amazon SageMaker AI Canvas

Amazon SageMaker AI Canvas vous aide à utiliser l'apprentissage automatique pour générer des prédictions sans avoir à écrire de code. Il fournit une interface visuelle sans code qui vous permet de préparer les données, de créer et de déployer des modèles de machine learning, rationalisant ainsi le cycle de vie de machine end-to-end learning dans un environnement unifié. Les complexités de la préparation des données, du développement de modèles, de la détection des biais, de l'explicabilité et de la surveillance sont résumées derrière une interface intuitive. Les utilisateurs n'ont pas besoin d'être des experts en SageMaker IA ou en opérations d'apprentissage automatique (MLOps) pour développer, opérationnaliser et surveiller des modèles avec SageMaker AI Canvas.

Avec SageMaker AI Canvas, la fonctionnalité RAG est fournie par le biais d'une fonction d'interrogation de documents sans code. Vous pouvez enrichir l'expérience de chat dans SageMaker AI Canvas en utilisant un index Amazon Kendra comme moteur de recherche d'entreprise sous-jacent. Pour plus d'informations, voir Extraire des informations à partir de documents à l'aide de requêtes de documents.

La connexion d' SageMaker AI Canvas à l'index Amazon Kendra nécessite une configuration unique. Dans le cadre de la configuration du domaine, un administrateur cloud peut choisir un ou plusieurs index Kendra que l'utilisateur peut interroger lorsqu'il interagit avec Canvas. SageMaker Pour savoir comment activer la fonctionnalité d'interrogation de documents, consultez Getting started with Amazon SageMaker AI Canvas.

SageMaker AI Canvas gère la communication sous-jacente entre Amazon Kendra et le modèle de base sélectionné. Pour plus d'informations sur les modèles de base pris en charge par SageMaker AI Canvas, voir Modèles de base d'IA générative dans SageMaker AI Canvas. Le schéma suivant montre comment fonctionne la fonctionnalité de requête de documents une fois que l'administrateur du cloud a connecté SageMaker AI Canvas à un index Amazon Kendra.



Le schéma suivant illustre le flux de travail suivant :

1. L'utilisateur lance une nouvelle discussion dans SageMaker AI Canvas, active l'option Query documents, sélectionne l'index cible, puis soumet une question.
2. SageMaker AI Canvas utilise la requête pour rechercher les données pertinentes dans l'index Amazon Kendra.
3. SageMaker AI Canvas extrait les données et leurs sources à partir de l'index Amazon Kendra.
4. SageMaker AI Canvas met à jour l'invite pour inclure le contexte extrait de l'index Amazon Kendra et soumet l'invite au modèle de base.
5. Le modèle de base utilise la question initiale et le contexte extrait pour générer une réponse.
6. SageMaker AI Canvas fournit la réponse générée à l'utilisateur. Il inclut des références aux sources de données, telles que les documents, qui ont été utilisées pour générer la réponse.

Architectures de génération augmentée à récupération personnalisée sur AWS

La section précédente décrit comment utiliser un RAG (Retrieval Augmented Generation) entièrement géré Service AWS . Cependant, certains cas d'utilisation nécessitent un contrôle accru sur les composants du système, tels que le retriever ou le LLM (également appelé générateur). Par exemple, vous pourriez avoir besoin de flexibilité pour choisir votre propre base de données vectorielle ou accéder à une source de données non prise en charge. Pour ces cas d'utilisation, vous pouvez créer une architecture RAG personnalisée.

Cette section contient les rubriques suivantes :

- Retrievers pour les flux de travail RAG
- Générateurs pour les flux de travail RAG

Pour plus d'informations sur le choix entre les options de récupération et de génération dans cette section, consultez Choix d'une option de génération augmentée de récupération sur AWS ce guide.

Retrievers pour les flux de travail RAG

Cette section explique comment créer un retriever. Vous pouvez utiliser une solution de recherche sémantique entièrement gérée, telle qu'Amazon Kendra, ou créer une recherche sémantique personnalisée à l'aide AWS d'une base de données vectorielle.

Avant de passer en revue les options du récupérateur, assurez-vous de bien comprendre les trois étapes du processus de recherche vectorielle :

1. Vous séparez les documents qui doivent être indexés en parties plus petites. C'est ce qu'on appelle le découpage.
2. Vous utilisez un processus appelé incorporation pour convertir chaque segment en un vecteur mathématique. Ensuite, vous indexez chaque vecteur dans une base de données vectorielle. L'approche que vous utilisez pour indexer les documents influence la rapidité et la précision de la recherche. L'approche d'indexation dépend de la base de données vectorielle et des options de configuration qu'elle fournit.
3. Vous convertissez la requête utilisateur en vecteur en utilisant le même processus. Le récupérateur recherche dans la base de données vectorielle des vecteurs similaires au vecteur

de requête de l'utilisateur. La similarité est calculée à l'aide de mesures telles que la distance euclidienne, la distance en cosinus ou le produit ponctuel.

Ce guide explique comment utiliser les services suivants Services AWS ou des services tiers pour créer une couche de récupération personnalisée sur AWS :

- Amazon Kendra
- Amazon OpenSearch Service
- Amazon Aurora, PostgreSQL et pgvector
- Amazon Neptune Analytics
- Amazon MemoryDB
- Amazon DocumentDB
- Pinecone
- MongoDB Atlas
- Weaviate

Amazon Kendra

Amazon Kendra est un service de recherche intelligent entièrement géré qui utilise le traitement du langage naturel et des algorithmes d'apprentissage automatique avancés pour renvoyer des réponses spécifiques aux questions de recherche à partir de vos données. Amazon Kendra vous permet d'ingérer directement des documents provenant de sources multiples et d'interroger les documents une fois qu'ils ont été correctement synchronisés. Le processus de synchronisation crée l'infrastructure nécessaire pour créer une recherche vectorielle sur le document ingéré. Par conséquent, Amazon Kendra ne nécessite pas les trois étapes traditionnelles du processus de recherche vectorielle. Après la synchronisation initiale, vous pouvez utiliser un calendrier défini pour gérer l'ingestion continue.

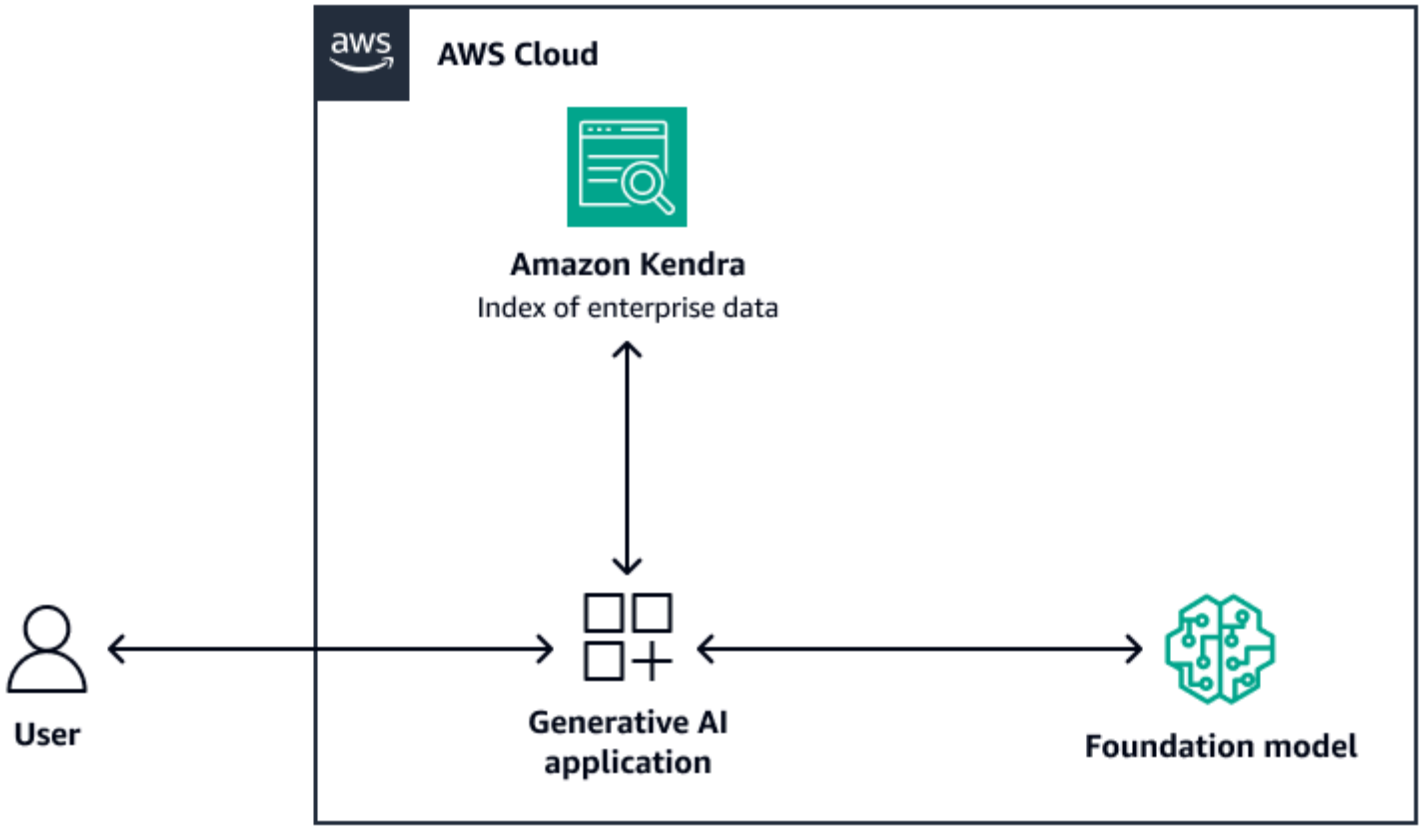
Les avantages de l'utilisation d'Amazon Kendra pour RAG sont les suivants :

- Il n'est pas nécessaire de gérer une base de données vectorielle car Amazon Kendra gère l'intégralité du processus de recherche vectorielle.
- Amazon Kendra contient des connecteurs prédéfinis pour les sources de données les plus courantes, telles que les bases de données, les robots d'exploration de sites Web, les

compartiments Amazon S3, Microsoft SharePoint les instances et les instances. Atlassian Confluence Des connecteurs développés par AWS des partenaires sont disponibles, tels que des connecteurs pour Box etGitLab.

- Amazon Kendra fournit un filtrage par liste de contrôle d'accès (ACL) qui renvoie uniquement les documents auxquels l'utilisateur final a accès.
- Amazon Kendra peut améliorer les réponses en fonction des métadonnées, telles que la date ou le référentiel des sources.

L'image suivante montre un exemple d'architecture qui utilise Amazon Kendra comme couche de récupération du système RAG. Pour plus d'informations, consultez Création rapide d'applications d'IA générative de haute précision sur des données d'entreprise à l'aide d'Amazon Kendra LangChain et de grands modèles linguistiquesAWS (article de blog).



Pour le modèle de base, vous pouvez utiliser Amazon Bedrock ou un LLM déployé via Amazon SageMaker AI. JumpStart Vous pouvez utiliser AWS Lambda with LangChainpour orchestrer le flux entre l'utilisateur, Amazon Kendra, et le LLM. Pour créer un système RAG qui utilise Amazon KendraLangChain, etc., consultez le LLMs référentiel Amazon LangChain Kendra Extensions. GitHub

Amazon OpenSearch Service

Amazon OpenSearch Service fournit des algorithmes ML intégrés pour la recherche des voisins les plus proches (k-NN) afin d'effectuer une recherche vectorielle. OpenSearch Le service fournit également un moteur vectoriel pour Amazon EMR Serverless. Vous pouvez utiliser ce moteur vectoriel pour créer un système RAG doté de capacités de stockage et de recherche vectorielles évolutives et performantes. Pour plus d'informations sur la création d'un système RAG à l'aide de OpenSearch Serverless, consultez Créer des flux de travail RAG évolutifs et sans serveur avec un moteur vectoriel pour les modèles Amazon Serverless OpenSearch et Amazon Bedrock Claude (article de blog).AWS

Les avantages de l'utilisation de OpenSearch Service pour la recherche vectorielle sont les suivants :

- Il fournit un contrôle complet sur la base de données vectorielle, y compris la création d'une recherche vectorielle évolutive à l'aide de OpenSearch Serverless.
- Il permet de contrôler la stratégie de segmentation.
- Il utilise les algorithmes du voisin le plus proche (ANN) approximatifs issus des bibliothèques Non-Metric Space Library (NMSLIB), Faiss et Apache Lucene pour effectuer une recherche K-nn. Vous pouvez modifier l'algorithme en fonction du cas d'utilisation. Pour plus d'informations sur les options de personnalisation de la recherche vectorielle via OpenSearch Service, consultez les fonctionnalités de la base de données vectorielle Amazon OpenSearch Service expliquées (article de AWS blog).
- OpenSearch Serverless s'intègre aux bases de connaissances Amazon Bedrock sous forme d'index vectoriel.

Amazon Aurora, PostgreSQL et pgvector

Amazon Aurora PostgreSQL Compatible Edition est un moteur de base de données relationnelle entièrement géré qui vous aide à configurer, exploiter et dimensionner les déploiements PostgreSQL. pgvector est une extension PostgreSQL open source qui fournit des fonctionnalités de recherche de similarité vectorielle. Cette extension est disponible à la fois pour la compatibilité avec Aurora PostgreSQL et pour Amazon Relational Database Service (Amazon RDS) pour PostgreSQL. Pour plus d'informations sur la création d'un système basé sur RAG qui utilise la compatibilité avec Aurora PostgreSQL et pgvector, consultez les articles de blog suivants : AWS

- Création d'une recherche basée sur l'IA dans PostgreSQL à l'aide d'Amazon AI et de pgvector SageMaker

- Tirez parti de pgvector et Amazon Aurora PostgreSQL pour le traitement du langage naturel, les chatbots et l'analyse des sentiments

Les avantages de l'utilisation de pgvector et de la compatibilité avec Aurora PostgreSQL sont les suivants :

- Il prend en charge la recherche exacte et approximative du voisin le plus proche. Il prend également en charge les métriques de similarité suivantes : distance L2, produit intérieur et distance en cosinus.
- Il prend en charge les fichiers inversés avec compression plate (IVFFlat) et l'indexation Hierarchical Navigable Small Worlds (HNSW).
- Vous pouvez combiner la recherche vectorielle avec des requêtes portant sur des données spécifiques à un domaine disponibles dans la même instance de PostgreSQL.
- La compatibilité avec Aurora PostgreSQL est optimisée I/O et fournit une mise en cache hiérarchisée. Pour les charges de travail qui dépassent la mémoire d'instance disponible, pgvector peut augmenter le nombre de requêtes par seconde pour la recherche vectorielle jusqu'à 8 fois.

Amazon Neptune Analytics

Amazon Neptune Analytics est un moteur de base de données graphique optimisé pour la mémoire à des fins d'analyse. Il prend en charge une bibliothèque d'algorithmes d'analyse de graphes optimisés, des requêtes graphiques à faible latence et des fonctionnalités de recherche vectorielle dans le cadre des traversées de graphes. Il dispose également d'une recherche de similarité vectorielle intégrée. Il fournit un point de terminaison pour créer un graphe, charger des données, invoquer des requêtes et effectuer une recherche de similarité vectorielle. Pour plus d'informations sur la création d'un système basé sur RAG qui utilise Neptune Analytics, consultez Utilisation de graphes de connaissances pour créer des applications GraphRag avec Amazon Bedrock et Amazon Neptune AWS (article de blog).

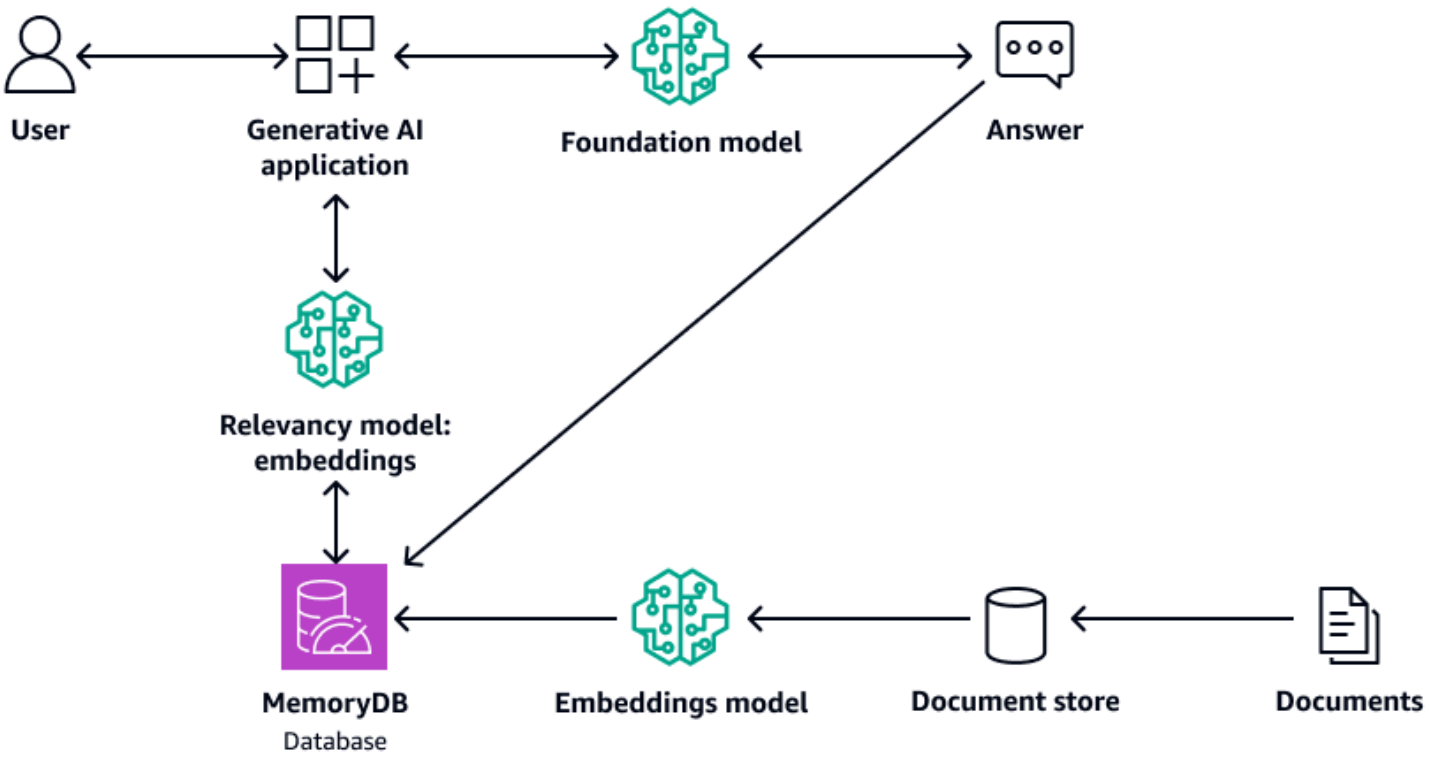
Les avantages de l'utilisation de Neptune Analytics sont les suivants :

- Vous pouvez stocker et rechercher des intégrations dans des requêtes graphiques.
- Si vous intégrez Neptune AnalyticsLangChain, cette architecture prend en charge les requêtes graphiques en langage naturel.
- Cette architecture stocke de grands ensembles de données graphiques en mémoire.

Amazon MemoryDB

Amazon MemoryDB est un service de base de données en mémoire durable qui fournit des performances ultrarapides. Toutes vos données sont stockées en mémoire, ce qui permet une lecture en microsecondes, une latence d'écriture d'un chiffre en millisecondes et un débit élevé. La recherche vectorielle pour MemoryDB étend les fonctionnalités de MemoryDB et peut être utilisée conjointement avec les fonctionnalités MemoryDB existantes. Pour plus d'informations, consultez le référentiel de réponses aux questions avec LLM et RAG sur. GitHub

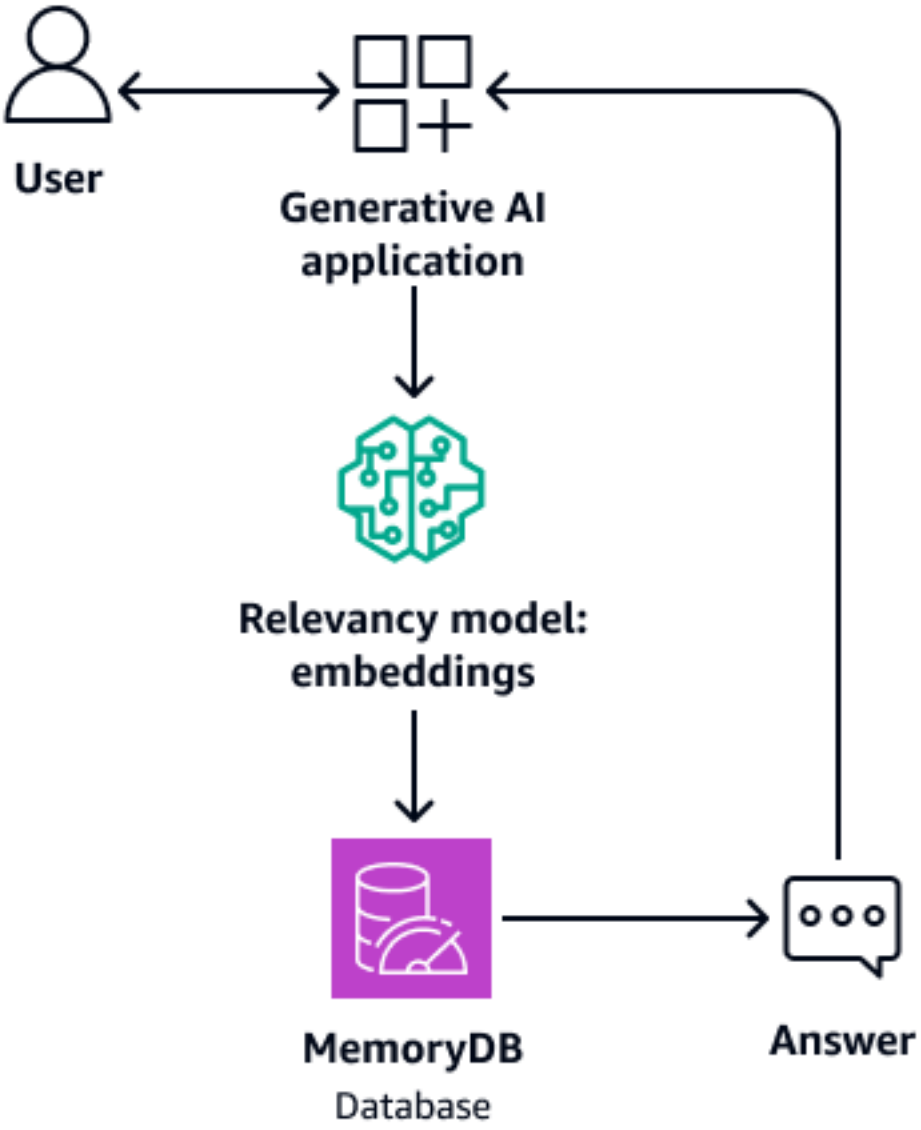
Le schéma suivant montre un exemple d'architecture qui utilise MemoryDB comme base de données vectorielle.



Les avantages de l'utilisation de MemoryDB sont les suivants :

- Il prend en charge les algorithmes d'indexation Flat et HNSW. Pour plus d'informations, consultez la section La recherche vectorielle pour Amazon MemoryDB est désormais disponible pour tous sur le AWS blog d'actualités
- Il peut également servir de mémoire tampon pour le modèle de base. Cela signifie que les questions auxquelles vous avez déjà répondu sont extraites de la mémoire tampon au lieu de

recommencer le processus de récupération et de génération. Le schéma suivant illustre ce processus.

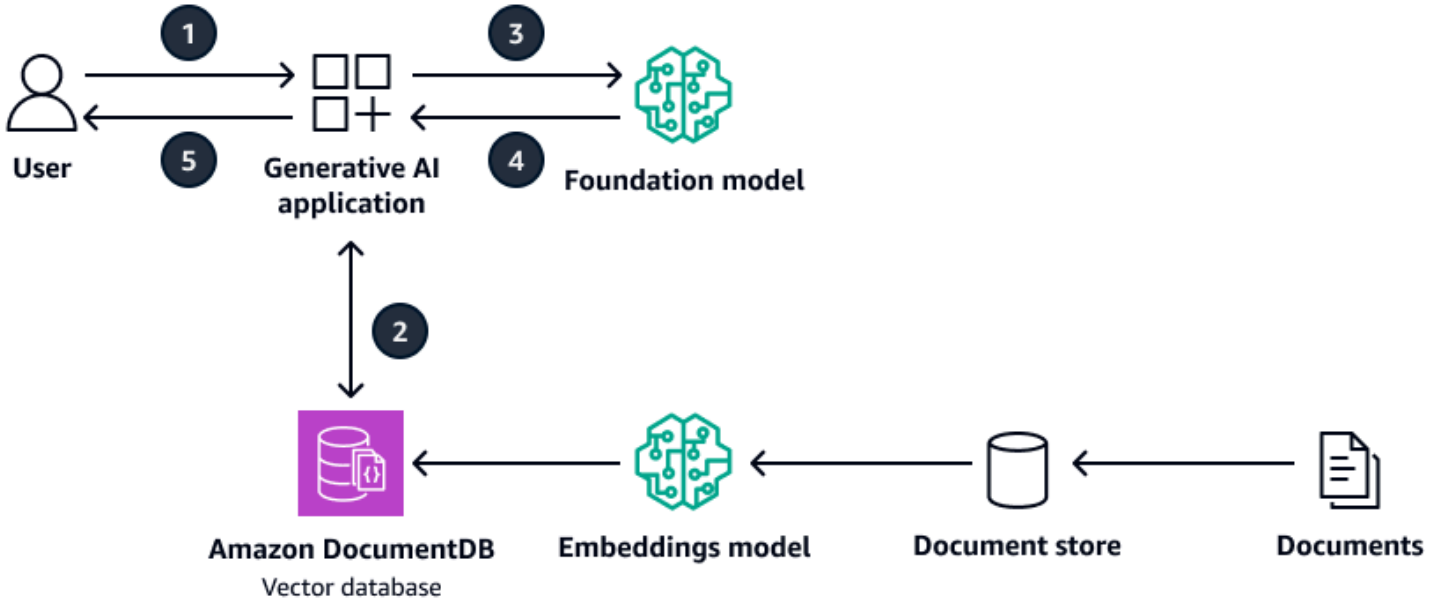


- Comme elle utilise une base de données en mémoire, cette architecture fournit un temps de requête à un chiffre en millisecondes pour la recherche sémantique.
- Il fournit jusqu'à 33 000 requêtes par seconde avec un taux de rappel de 95 à 99 % et 26 500 requêtes par seconde avec un taux de rappel supérieur à 99 %. Pour plus d'informations, consultez la vidéo AWS re:Invent 2023 - Recherche vectorielle à très faible latence pour Amazon MemoryDB sur. YouTube

Amazon DocumentDB

Amazon DocumentDB (compatible avec MongoDB) est un service de base de données rapide, fiable et entièrement géré. Il facilite la configuration, l'exploitation et le dimensionnement de bases de données MongoDB compatibles dans le cloud. La recherche vectorielle pour Amazon DocumentDB associe la flexibilité et la riche capacité d'interrogation d'une base de données de documents basée sur JSON à la puissance de la recherche vectorielle. Pour plus d'informations, consultez le référentiel de réponses aux questions avec LLM et RAG sur. GitHub

Le schéma suivant montre un exemple d'architecture qui utilise Amazon DocumentDB comme base de données vectorielle.



Le schéma suivant illustre le flux de travail suivant :

1. L'utilisateur soumet une requête à l'application d'IA générative.
2. L'application d'IA générative effectue une recherche de similarité dans la base de données vectorielle Amazon DocumentDB et extrait les extraits de documents pertinents.
3. L'application d'IA générative met à jour la requête de l'utilisateur avec le contexte extrait et soumet l'invite au modèle de base cible.
4. Le modèle de base utilise le contexte pour générer une réponse à la question de l'utilisateur et renvoie la réponse.
5. L'application d'IA générative renvoie la réponse à l'utilisateur.

Les avantages de l'utilisation d'Amazon DocumentDB sont les suivants :

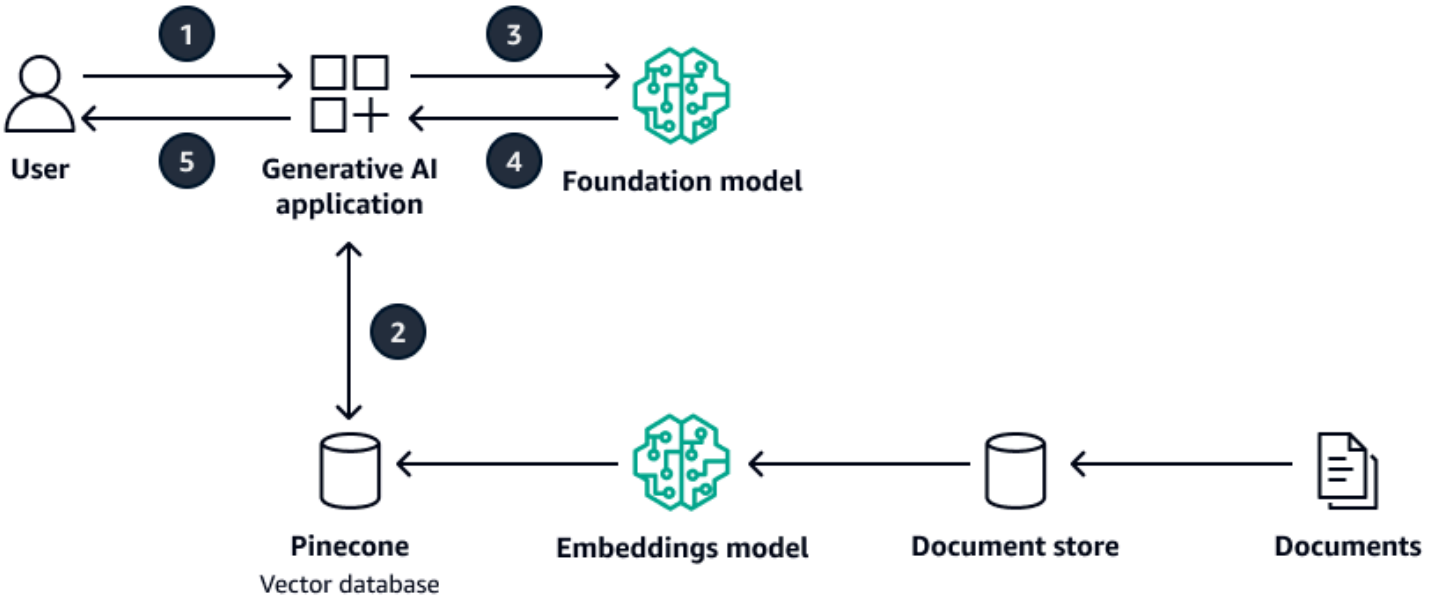
- Il prend en charge à la fois les méthodes HNSW et les IVFFlat méthodes d'indexation.
- Il prend en charge jusqu'à 2 000 dimensions dans les données vectorielles et prend en charge les mesures de distance entre les produits euclidiens, cosinusoïdaux et points.
- Il fournit des temps de réponse en millisecondes.

Pinecone

Pineconeest une base de données vectorielle entièrement gérée qui vous permet d'ajouter la recherche vectorielle aux applications de production. Il est disponible via le AWS Marketplace. La facturation est basée sur l'utilisation, et les frais sont calculés en multipliant le prix des capsules par le nombre de capsules. Pour plus d'informations sur la création d'un système basé sur RAG qui utilisePinecone, consultez les articles de AWS blog suivants :

- Atténuez les hallucinations grâce à RAG en utilisant la base de données Pinecone vectorielle et Llama-2 d'Amazon AI SageMaker JumpStart
- Utilisez Amazon SageMaker AI Studio pour créer une solution de réponse aux questions RAG avec Llama 2 et Pinecone pour une expérimentation rapide LangChain

Le schéma suivant montre un exemple d'architecture utilisé Pinecone comme base de données vectorielle.



Le schéma suivant illustre le flux de travail suivant :

1. L'utilisateur soumet une requête à l'application d'IA générative.
2. L'application d'IA générative effectue une recherche de similarité dans la base de données Pinecone vectorielle et récupère les extraits de documents pertinents.
3. L'application d'IA générative met à jour la requête de l'utilisateur avec le contexte extrait et soumet l'invite au modèle de base cible.
4. Le modèle de base utilise le contexte pour générer une réponse à la question de l'utilisateur et renvoie la réponse.
5. L'application d'IA générative renvoie la réponse à l'utilisateur.

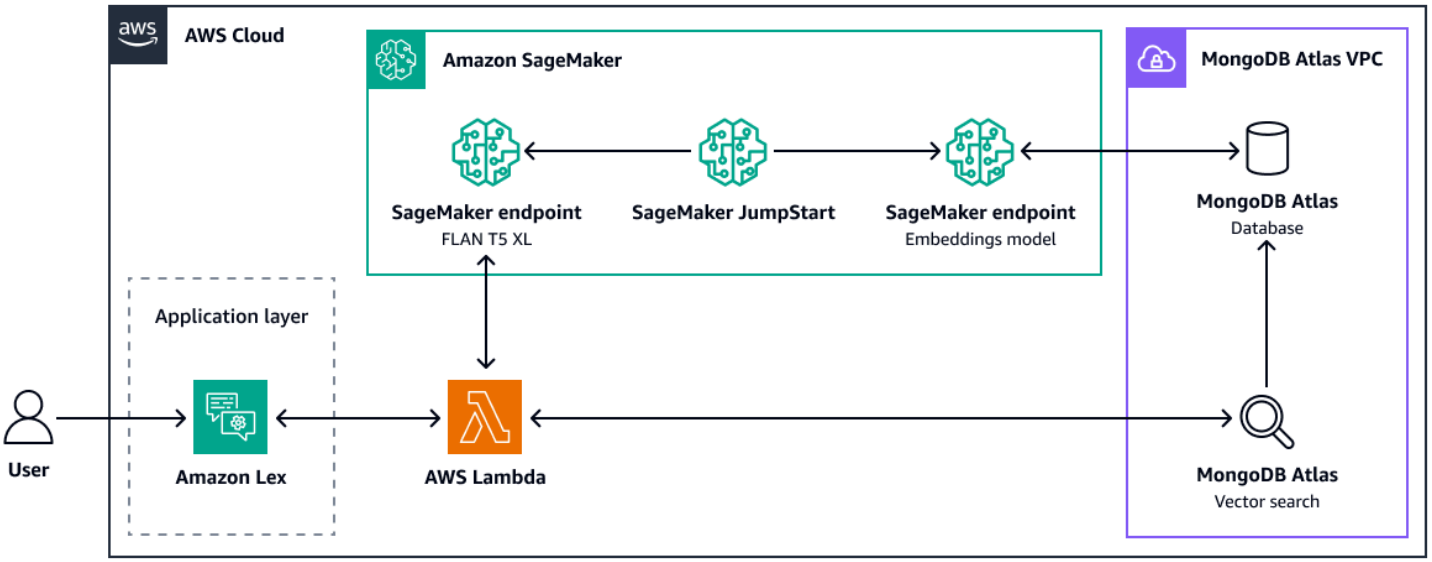
Les avantages de l'utilisation sont les Pinecone suivants :

- Il s'agit d'une base de données vectorielle entièrement gérée qui vous évite de devoir gérer votre propre infrastructure.
- Il fournit des fonctionnalités supplémentaires de filtrage, de mise à jour d'index en direct et de renforcement des mots clés (recherche hybride).

MongoDB Atlas

MongoDB Atlasest une base de données cloud entièrement gérée qui gère toute la complexité du déploiement et de la gestion de vos déploiements sur AWS. Vous pouvez utiliser la recherche vectorielle MongoDB Atlas pour stocker des intégrations vectorielles dans votre base de données. MongoDB Les bases de connaissances Amazon Bedrock prennent en MongoDB Atlas charge le stockage vectoriel. Pour plus d'informations, consultez la section Commencer à intégrer la base de connaissances Amazon Bedrock dans la MongoDB documentation.

Pour plus d'informations sur l'utilisation de la recherche MongoDB Atlas vectorielle pour RAG, consultez Retrieval-Augmented Generation with, LangChain Amazon SageMaker AI JumpStart et MongoDB Atlas Semantic Search (article de blog).AWS Le schéma suivant montre l'architecture de la solution détaillée dans ce billet de blog.



Les avantages de la recherche MongoDB Atlas vectorielle sont les suivants :

- Vous pouvez utiliser votre implémentation existante MongoDB Atlas pour stocker et rechercher des intégrations vectorielles.
- Vous pouvez utiliser l'API MongoDB Query pour interroger les intégrations vectorielles.
- Vous pouvez redimensionner indépendamment la recherche vectorielle et la base de données.
- Les intégrations vectorielles sont stockées à proximité des données sources (documents), ce qui améliore les performances d'indexation.

Weaviate

Weaviateest une base de données vectorielle open source populaire à faible latence qui prend en charge les types de médias multimodaux, tels que le texte et les images. La base de données stocke à la fois des objets et des vecteurs, ce qui combine la recherche vectorielle avec le filtrage structuré. Pour plus d'informations sur l'utilisation d'Amazon Bedrock Weaviate et sur la création d'un flux de travail RAG, consultez Créer des solutions d'IA générative adaptées aux entreprises avec les modèles de base de données Cohere dans Amazon Bedrock et une base de données Weaviate vectorielle sur (article de blog). AWS MarketplaceAWS

Les avantages de l'utilisation sont les Weaviate suivants :

- Il est open source et soutenu par une communauté solide.
- Il est conçu pour la recherche hybride (vecteurs et mots clés).

- Vous pouvez le déployer en AWS tant qu'offre de logiciel géré en tant que service (SaaS) ou en tant que cluster Kubernetes.

Générateurs pour les flux de travail RAG

Les grands modèles de langage (LLMs) sont de très grands modèles d'apprentissage profond qui sont préentraînés sur de grandes quantités de données. Ils sont incroyablement flexibles. LLMs peut effectuer diverses tâches, telles que répondre à des questions, résumer des documents, traduire des langues et compléter des phrases. Ils ont le potentiel de perturber la création de contenu et la façon dont les utilisateurs utilisent les moteurs de recherche et les assistants virtuels. Bien que cela ne soit pas parfait, faites LLMs preuve d'une capacité remarquable à faire des prédictions sur la base d'une invite ou d'un nombre relativement restreint d'entrées.

LLMs sont un élément essentiel d'une solution RAG. Pour les architectures RAG personnalisées, il en existe deux Services AWS qui constituent les principales options :

- Amazon Bedrock est un service entièrement géré qui met LLMs à votre disposition les principales sociétés d'IA et Amazon via une API unifiée.
- Amazon SageMaker AI JumpStart est un hub de machine learning qui propose des modèles de base, des algorithmes intégrés et des solutions de machine learning prédéfinies. Grâce à SageMaker l'IA JumpStart, vous pouvez accéder à des modèles préentraînés, y compris des modèles de base. Vous pouvez également utiliser vos propres données pour affiner les modèles préentraînés.

Amazon Bedrock

Amazon Bedrock propose des modèles de pointe deAnthropic,Stability AI,Meta,Cohere, AI21 LabsMistral AI, et Amazon. Pour obtenir la liste complète, consultez la section Modèles de fondation pris en charge dans Amazon Bedrock. Amazon Bedrock vous permet également de personnaliser des modèles avec vos propres données.

Vous pouvez évaluer les performances du modèle afin de déterminer lequel est le mieux adapté à votre cas d'utilisation de RAG. Vous pouvez tester les derniers modèles et également tester pour voir quelles capacités et fonctionnalités fournissent les meilleurs résultats au meilleur prix. Le modèle Anthropic Claude Sonnet est un choix courant pour les applications RAG car il excelle dans un large éventail de tâches et offre un haut degré de fiabilité et de prévisibilité.

SageMaker AI JumpStart

SageMaker L'IA JumpStart fournit des modèles open source préentraînés pour un large éventail de types de problèmes. Vous pouvez entraîner et peaufiner progressivement ces modèles avant leur déploiement. Vous pouvez accéder aux modèles préentraînés, aux modèles de solutions et aux exemples via la page JumpStart d'accueil SageMaker AI d'Amazon SageMaker AI Studio ou utiliser le SDK SageMaker AI Python.

SageMaker L'IA JumpStart propose des modèles de state-of-the-art base pour des cas d'utilisation tels que la rédaction de contenu, la génération de code, la réponse aux questions, la rédaction, la synthèse, la classification, la récupération d'informations, etc. Utilisez des modèles de JumpStart base pour créer vos propres solutions d'IA générative et intégrez des solutions personnalisées avec des fonctionnalités d' SageMaker IA supplémentaires. Pour plus d'informations, consultez Getting started with Amazon SageMaker AI JumpStart.

SageMaker L'IA JumpStart intègre et gère des modèles de base accessibles au public auxquels vous pouvez accéder, personnaliser et intégrer à vos cycles de vie du machine learning. Pour plus d'informations, consultez la section Modèles de fondation accessibles au public. SageMaker L'IA inclut JumpStart également des modèles de base propriétaires provenant de fournisseurs tiers. Pour plus d'informations, consultez la section Modèles de fondation propriétaires.

Choix d'une option de génération augmentée de récupération sur AWS

Les sections Options RAG entièrement gérées et Architectures RAG personnalisées de ce guide décrivent différentes approches pour créer une solution de recherche basée sur RAG. AWS Cette section explique comment sélectionner l'une de ces options en fonction de votre cas d'utilisation. Dans certains cas, plusieurs options peuvent fonctionner. Dans ce scénario, le choix dépend de la facilité de mise en œuvre, des compétences disponibles dans votre organisation et des politiques et normes de votre entreprise.

Nous vous recommandons de prendre en compte les options RAG entièrement gérées et personnalisées dans l'ordre suivant et de choisir la première option qui correspond à votre cas d'utilisation :

1. Utilisez Amazon Q Business sauf si :
   - Ce service n'est pas disponible dans votre Région AWS région et vos données ne peuvent pas être déplacées vers une région où elles sont disponibles
   - Vous avez une raison précise de personnaliser le flux de travail RAG
   - Vous souhaitez utiliser une base de données vectorielle existante ou un LLM spécifique
2. Utilisez les bases de connaissances pour Amazon Bedrock sauf si :
   - Vous avez une base de données vectorielle qui n'est pas prise en charge
   - Vous avez une raison précise de personnaliser le flux de travail RAG
3. Combinez Amazon Kendra avec le générateur de votre choix, sauf si :
   - Vous souhaitez choisir votre propre base de données vectorielles
   - Vous souhaitez personnaliser la stratégie de segmentation
4. Si vous souhaitez mieux contrôler le récupérateur et sélectionner votre propre base de données vectorielles :
   - Si vous ne disposez pas d'une base de données vectorielle existante et que vous n'avez pas besoin d'une faible latence ou de requêtes graphiques, pensez à utiliser Amazon OpenSearch Service.
   - Si vous possédez déjà une base de données PostgreSQL vectorielle, pensez à utiliser l'option Amazon Aurora PostgreSQL and pgvector.

- Si vous avez besoin d'une faible latence, envisagez une option en mémoire, telle qu'Amazon MemoryDB ou Amazon DocumentDB.
- Si vous souhaitez associer la recherche vectorielle à une requête graphique, pensez à Amazon Neptune Analytics.
- Si vous utilisez déjà une base de données vectorielle tierce ou si vous y trouvez un avantage spécifique, considérez PineconeMongoDB Atlas, et Weaviate.

5. Si vous souhaitez choisir un LLM :
   - Si vous utilisez Amazon Q Business, vous ne pouvez pas choisir le LLM.
   - Si vous utilisez Amazon Bedrock, vous pouvez choisir l'un des modèles de fondation pris en charge.
   - Si vous utilisez Amazon Kendra ou une base de données vectorielle personnalisée, vous pouvez utiliser l'un des générateurs décrits dans ce guide ou utiliser un LLM personnalisé.

   > **Note**
   >
   > Vous pouvez également utiliser vos documents personnalisés pour affiner un LLM existant afin d'augmenter la précision de ses réponses. Pour plus d’informations, consultez Comparaison entre RAG et réglage fin dans ce guide.

6. Si vous souhaitez utiliser une implémentation existante d'Amazon SageMaker AI Canvas ou si vous souhaitez comparer des réponses RAG différentes LLMs, pensez à Amazon SageMaker AI Canvas.

Conclusion

Ce guide décrit les différentes options pour créer un système RAG (Retrieval Augmented Generation). AWS Vous pouvez commencer par des services entièrement gérés, tels que les bases de connaissances Amazon Q Business et Amazon Bedrock. Si vous souhaitez mieux contrôler le flux de travail RAG, vous pouvez choisir un récupérateur personnalisé. Pour un générateur, vous pouvez utiliser une API pour appeler un LLM pris en charge dans Amazon Bedrock, ou vous pouvez déployer votre propre LLM à l'aide d'Amazon AI. SageMaker JumpStart Consultez les recommandations de la section Choisir une option RAG pour déterminer l'option la mieux adaptée à votre cas d'utilisation. Après avoir sélectionné la meilleure option pour votre cas d'utilisation, utilisez les références fournies dans ce guide pour commencer à créer votre application basée sur RAG.

Historique du document

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

Modification	Description	Date
Publication initiale	—	28 octobre 2024

AWS Glossaire des directives prescriptives

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

Nombres

7 R

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

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

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

A

ABAC

Voir contrôle d'accès basé sur les attributs.

services abstraits

Consultez la section Services gérés.

ACIDE

Voir atomicité, consistance, isolation, durabilité.

migration active-active

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

migration active-passive

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

fonction d'agrégation

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

AI

Voir intelligence artificielle.

AIOps

Voir les opérations d'intelligence artificielle.

anonymisation

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

anti-motif

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

contrôle des applications

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

portefeuille d'applications

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

intelligence artificielle (IA)

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

opérations d'intelligence artificielle (AIOps)

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

chiffrement asymétrique

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

atomicité, cohérence, isolement, durabilité (ACID)

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

contrôle d'accès par attributs (ABAC)

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

source de données faisant autorité

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

Zone de disponibilité

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

AWS Cadre d'adoption du cloud (AWS CAF)

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

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

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

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

B

mauvais bot

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

BCP

Consultez la section Planification de la continuité des activités.

graphique de comportement

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

système de poids fort

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

classification binaire

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

filtre de Bloom

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

déploiement bleu/vert

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

bot

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

botnet

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

branche

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

accès par brise-vitre

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

stratégie existante (brownfield)

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

cache de tampon

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

capacité métier

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

planification de la continuité des activités (BCP)

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

C

CAF

Voir le cadre d'adoption du AWS cloud.

déploiement de Canary

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

CCo E

Voir le Centre d'excellence du cloud.

CDC

Voir capture des données de modification.

capture des données de modification (CDC)

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

ingénierie du chaos

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

CI/CD

Découvrez l'intégration continue et la livraison continue.

classification

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

chiffrement côté client

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

Centre d'excellence du cloud (CCoE)

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

cloud computing

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

modèle d'exploitation du cloud

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

étapes d'adoption du cloud

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

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

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

CMDB

Consultez la base de données de gestion des configurations.

référentiel de code

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

cache passif

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

données gelées

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

vision par ordinateur (CV)

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

dérive de configuration

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

base de données de gestion des configurations (CMDB)

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

pack de conformité

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

intégration continue et livraison continue (CI/CD)

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

CV

Voir vision par ordinateur.

D

données au repos

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

classification des données

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

dérive des données

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

données en transit

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

maillage de données

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

minimisation des données

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

périmètre de données

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

prétraitement des données

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

provenance des données

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

sujet des données

Personne dont les données sont collectées et traitées.

entrepôt des données

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

langage de définition de base de données (DDL)

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

langage de manipulation de base de données (DML)

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

DDL

Voir langage de définition de base de données.

ensemble profond

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

deep learning

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

defense-in-depth

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

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

administrateur délégué

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

déploiement

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

environnement de développement

Voir environnement.

contrôle de détection

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

cartographie de la chaîne de valeur du développement (DVSM)

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

jumeau numérique

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

tableau des dimensions

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

catastrophe

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

reprise après sinistre (DR)

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

DML

Voir langage de manipulation de base de données.

conception axée sur le domaine

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

DR

Voir reprise après sinistre.

détection de dérive

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

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

DVSM

Voir la cartographie de la chaîne de valeur du développement.

E

EDA

Voir analyse exploratoire des données.

EDI

Voir échange de données informatisé.

informatique de périphérie

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

échange de données informatisé (EDI)

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

chiffrement

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

clé de chiffrement

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

endianisme

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

point de terminaison

Voir point de terminaison de service.

service de point de terminaison

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

planification des ressources d'entreprise (ERP)

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

chiffrement d'enveloppe

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

environment

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

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

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

époppée

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

ERP

Voir Planification des ressources d'entreprise.

analyse exploratoire des données (EDA)

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

F

tableau des faits

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

échouer rapidement

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

limite d'isolation des défauts

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

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

branche de fonctionnalités

Voir succursale.

fonctionnalités

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

importance des fonctionnalités

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

transformation de fonctionnalité

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

invitation en quelques coups

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

FGAC

Découvrez le contrôle d'accès détaillé.

contrôle d'accès détaillé (FGAC)

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

migration instantanée (flash-cut)

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

FM

Voir le modèle de fondation.

modèle de fondation (FM)

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

G

IA générative

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

blocage géographique

Voir les restrictions géographiques.

restrictions géographiques (blocage géographique)

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

Flux de travail Gitflow

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

image dorée

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

stratégie inédite

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

barrière de protection

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

H

HA

Découvrez la haute disponibilité.

migration de base de données hétérogène

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

haute disponibilité (HA)

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

modernisation des historiens

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

données de rétention

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

migration de base de données homogène

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

données chaudes

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

correctif

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

période de soins intensifs

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

I

IaC

Considérez l'infrastructure comme un code.

politique basée sur l’identité

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

application inactive

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

IIo T

Voir Internet industriel des objets.

infrastructure immuable

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

VPC entrant (d'entrée)

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

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

migration incrémentielle

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

Industry 4.0

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

infrastructure

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

infrastructure en tant que code (IaC)

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

Internet industriel des objets (IIoT)

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

VPC d'inspection

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

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

Internet des objets (IoT)

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

interprétabilité

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

IoT

Voir Internet des objets.

Bibliothèque d'informations informatiques (ITIL)

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

gestion des services informatiques (ITSM)

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

ITIL

Consultez la bibliothèque d'informations informatiques.

ITSM

Voir Gestion des services informatiques.

L

contrôle d'accès basé sur des étiquettes (LBAC)

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

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

zone de destination

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

grand modèle de langage (LLM)

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

migration de grande envergure

Migration de 300 serveurs ou plus.

LBAC

Voir contrôle d'accès basé sur des étiquettes.

principe de moindre privilège

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

lift and shift

Voir 7 Rs.

système de poids faible

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

LLM

Voir le grand modèle de langage.

environnements inférieurs

Voir environnement.

M

machine learning (ML)

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

branche principale

Voir succursale.

malware

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

services gérés

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

système d'exécution de la fabrication (MES)

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

MAP

Voir Migration Acceleration Program.

mécanisme

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

compte membre

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

MAILLES

Voir le système d'exécution de la fabrication.

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

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

microservice

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

architecture de microservices

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

Programme d'accélération des migrations (MAP)

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

migration à grande échelle

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

usine de migration

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

métadonnées de migration

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

modèle de migration

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

Évaluation du portefeuille de migration (MPA)

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

Évaluation de la préparation à la migration (MRA)

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

stratégie de migration

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

ML

Voir apprentissage automatique.

modernisation

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

évaluation de la préparation à la modernisation

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

applications monolithiques (monolithes)

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

MPA

Voir Évaluation du portefeuille de migration.

MQTT

Voir Message Queuing Telemetry Transport.

classification multi-classes

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

infrastructure mutable

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

O

OAC

Voir Contrôle d'accès à l'origine.

OAI

Voir l'identité d'accès à l'origine.

OCM

Voir gestion du changement organisationnel.

migration hors ligne

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

OI

Consultez la section Intégration des opérations.

OLA

Voir l'accord au niveau opérationnel.

migration en ligne

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

OPC-UA

Voir Open Process Communications - Architecture unifiée.

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

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

accord au niveau opérationnel (OLA)

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

examen de l'état de préparation opérationnelle (ORR)

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

technologie opérationnelle (OT)

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

intégration des opérations (OI)

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

journal de suivi d'organisation

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

gestion du changement organisationnel (OCM)

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

contrôle d'accès d'origine (OAC)

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

identité d'accès d'origine (OAI)

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

ORR

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

DE

Voir technologie opérationnelle.

VPC sortant (de sortie)

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

P

limite des autorisations

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

informations personnelles identifiables (PII)

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

PII

Voir les informations personnelles identifiables.

manuel stratégique

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

PLC

Voir contrôleur logique programmable.

PLM

Consultez la section Gestion du cycle de vie des produits.

policy

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

persistance polyglotte

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

évaluation du portefeuille

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

predicate

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

prédicat pushdown

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

contrôle préventif

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

principal

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

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

confidentialité dès la conception

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

zones hébergées privées

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

contrôle proactif

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

gestion du cycle de vie des produits (PLM)

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

environnement de production

Voir environnement.

contrôleur logique programmable (PLC)

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

chaînage rapide

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

pseudonymisation

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

publish/subscribe (pub/sub)

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

Q

plan de requête

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

régression du plan de requêtes

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

R

Matrice RACI

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

RAG

Voir Retrieval Augmented Generation.

rançongiciel

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

Matrice RASCI

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

RCAC

Voir contrôle d'accès aux lignes et aux colonnes.

réplica en lecture

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

réarchitecte

Voir 7 Rs.

objectif de point de récupération (RPO)

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

objectif de temps de récupération (RTO)

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

refactoriser

Voir 7 Rs.

Région

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

régression

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

réhéberger

Voir 7 Rs.

version

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

déplacer

Voir 7 Rs.

replateforme

Voir 7 Rs.

rachat

Voir 7 Rs.

résilience

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

politique basée sur les ressources

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

matrice responsable, redevable, consulté et informé (RACI)

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

contrôle réactif

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

retain

Voir 7 Rs.

se retirer

Voir 7 Rs.

Génération augmentée de récupération (RAG)

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

rotation

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

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

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

RPO

Voir l'objectif du point de récupération.

RTO

Voir l'objectif en matière de temps de rétablissement.

runbook

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

S

SAML 2.0

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

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

SCADA

Voir Contrôle de supervision et acquisition de données.

SCP

Voir la politique de contrôle des services.

secret

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

sécurité dès la conception

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

contrôle de sécurité

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

renforcement de la sécurité

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

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

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

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

automatisation des réponses de sécurité

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

chiffrement côté serveur

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

Politique de contrôle des services (SCP)

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

point de terminaison du service

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

contrat de niveau de service (SLA)

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

indicateur de niveau de service (SLI)

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

objectif de niveau de service (SLO)

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

modèle de responsabilité partagée

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

SIEM

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

point de défaillance unique (SPOF)

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

SLA

Voir le contrat de niveau de service.

SLI

Voir l'indicateur de niveau de service.

SLO

Voir l'objectif de niveau de service.

split-and-seed modèle

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

SPOF

Voir point de défaillance unique.

schéma en étoile

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

modèle de figuier étrangleur

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

sous-réseau

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

contrôle de supervision et acquisition de données (SCADA)

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

chiffrement symétrique

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

tests synthétiques

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

invite du système

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

T

tags

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

variable cible

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

liste de tâches

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

environnement de test

Voir environnement.

entraînement

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

passerelle de transit

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

flux de travail basé sur jonction

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

accès sécurisé

Accorder des autorisations à un service que vous spécifiez pour effectuer des tâches au sein de votre organisation AWS Organizations et dans ses comptes en votre nom. Le service de

confiance crée un rôle lié au service dans chaque compte, lorsque ce rôle est nécessaire, pour effectuer des tâches de gestion à votre place. Pour plus d'informations, consultez la section Utilisation AWS Organizations avec d'autres AWS services dans la AWS Organizations documentation.

réglage

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

équipe de deux pizzas

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

U

incertitude

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

tâches indifférenciées

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

environnements supérieurs

Voir environnement.

V

mise à vide

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

contrôle de version

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

Appairage de VPC

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

vulnérabilités

Défaut logiciel ou matériel qui compromet la sécurité du système.

W

cache actif

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

données chaudes

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

fonction de fenêtre

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

charge de travail

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

flux de travail

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

VER

Voir écrire une fois, lire plusieurs.

WQF

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

écrire une fois, lire plusieurs (WORM)

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

Z

exploit Zero-Day

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

vulnérabilité « jour zéro »

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

invite Zero-Shot

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

application zombie

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

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