Guide de mise en œuvre

Générateur d'applications d'IA générative sur AWS

Générateur d'applications d'IA générative sur AWS: Guide de mise en œuvre

Table of Contents

Cette solution facilite le développement, l'expérimentation rapide et le déploiement d'applications d'intelligence artificielle générative (IA)

Generative AI Application Builder sur AWS facilite le développement, l'expérimentation rapide et le déploiement d'applications d'intelligence artificielle générative (IA) sans nécessiter une expérience approfondie en matière d'IA. Cette solution AWS accélère le développement et rationalise l'expérimentation en vous aidant à :

- Ingérez des données et des documents spécifiques à votre entreprise
- Évaluer et comparer les performances de grands modèles linguistiques (LLMs)
- Exécutez des tâches et des flux de travail en plusieurs étapes avec des agents d'intelligence artificielle
- Créez rapidement des applications extensibles et déployez ces applications avec une architecture d'entreprise

Generative AI Application Builder sur AWS inclut des intégrations avec :

- LLMs disponible sur Amazon Bedrock
- LLMs que vous avez déployé sur Amazon SageMaker AI
- Bases de connaissances Amazon Bedrock pour la génération augmentée par extraction (RAG)
- Amazon Bedrock Guardrails va mettre en œuvre des mesures de protection et réduire les hallucinations
- Amazon Bedrock Agents pour créer des flux de travail agentiques capables d'orchestrer et de terminer les tâches
- Amazon Bedrock va AgentCore créer, déployer et gérer des agents d'IA prêts pour la production avec un support d'exécution étendu
- Serveurs MCP (Model Context Protocol) pour l'intégration des données et des outils d'entreprise

De plus, cette solution permet de se connecter au modèle de votre choix à l'aide de LangChain connecteurs. Ces connecteurs sont disponibles dans une fonction AWS Lambda déployée avec la solution. Vous pouvez commencer par l'assistant de déploiement sans code pour créer des

applications d'IA génératives pour la recherche conversationnelle, les chatbots générés par l'IA, la génération de texte et le résumé de texte.

Ce guide de mise en œuvre fournit une vue d'ensemble de la solution Generative AI Application Builder sur AWS, de son architecture de référence et de ses composants, des considérations relatives à la planification du déploiement et des étapes de configuration pour le déploiement de la solution sur le cloud Amazon Web Services (AWS).

Ce guide est destiné aux architectes de solutions, aux décideurs commerciaux, aux DevOps ingénieurs, aux data scientists et aux professionnels du cloud qui souhaitent implémenter Generative AI Application Builder sur AWS dans leur environnement.

Utilisez ce tableau de navigation pour trouver rapidement les réponses aux questions suivantes :

Si tu veux...	Lisez.
Connaissez le coût de fonctionnement de cette solution. Le coût estimé d'exécution de cette solution varie en fonction des composants que vous déployez et du nombre de requêtes. Le coût d'exécution du tableau de bord de déploiement avec les paramètres par défaut et 100 utilisateurs actifs dans la région de l'est des États-Unis (Virginie du Nord) pendant un mois est d'environ 20,12 USD par mois. Le coût d'un cas d'utilisation de texte déployé sans RAG pour 1 utilisateur professionnel effectuant 100 requêtes par jour avec le LLM est d'environ 12,39 USD par mois. Le coût d'un cas d'utilisation compatible avec RAG avec un indice Amazon Kendra prenant en charge 8 000 interactions par jour est d'environ 204,26 USD par mois, plus le coût de la base de connaissances.	Coût

Si tu veux...	Lisez.
Comprenez les considérations de sécurité liées à cette solution.	Sécurité
Sachez comment planifier les quotas pour cette solution.	Quotas
Découvrez quelles régions AWS prennent en charge cette solution.	Régions AWS prises en charge
Consultez ou téléchargez le CloudFormation modèle AWS inclus dans cette solution pour déployer automatiquement les ressources d'infrastructure (la « pile ») de cette solution.	CloudFormation Modèle AWS
Accédez au code source et utilisez éventuell ement l'AWS Cloud Development Kit (AWS CDK) pour déployer la solution.	GitHub référentiel

Fonctionnalités et avantages

La solution Generative AI Application Builder sur AWS fournit les fonctionnalités suivantes :

Expérimentation rapide

Cette solution permet aux utilisateurs d'expérimenter rapidement en supprimant les tâches fastidieuses nécessaires au déploiement de plusieurs instances avec différentes configurations et en comparant les résultats et les performances. Testez plusieurs configurations basées sur une ingénierie rapide LLMs, des bases de connaissances d'entreprise, des garde-corps, des agents d'intelligence artificielle et d'autres paramètres.

Choix et configurabilité

Avec des connecteurs prédéfinis pour une variété de modèles LLMs, tels que les modèles disponibles via Amazon Bedrock, cette solution vous donne la flexibilité de déployer le modèle de votre choix, ainsi que l'AWS et les principaux services FM que vous préférez. Vous pouvez également permettre aux agents Amazon Bedrock d'exécuter diverses tâches et flux de travail.

Générateur d'agents

Créez et déployez des agents d'IA prêts pour la production avec une gestion complète du cycle de vie. Configurez les instructions du système, intégrez des serveurs MCP (Model Context Protocol) pour les outils d'entreprise et l'accès aux données, et activez les capacités de mémoire pour la rétention du contexte dans les conversations. Les agents sont déployés sur Amazon Bedrock AgentCore avec un support d'exécution étendu et des réponses de streaming en temps réel.

Générateur de flux de travail

Orchestrez plusieurs agents Agent Builder dans des flux de travail complexes à l'aide de la délégation hiérarchique. Créez un agent superviseur qui sélectionne et coordonne de manière autonome les agents spécialisés d'Agent Builder pour gérer des tâches en plusieurs étapes. Configurez les descriptions des agents, les stratégies de délégation et la mémoire au niveau du flux de travail tout en réutilisant les déploiements d'Agent Builder existants.

Prêt pour la production

Conçue selon les principes de conception d'AWS Well-Architected, cette solution offre une sécurité et une évolutivité de niveau professionnel, ainsi qu'une haute disponibilité et une faible latence, garantissant une intégration fluide dans vos applications avec des normes de performance élevées.

Architecture modulaire extensible

Étendez les fonctionnalités de cette solution en intégrant vos projets existants ou en connectant nativement des services AWS supplémentaires. Comme il s'agit d'une application open source, vous pouvez utiliser la couche d' LangChain orchestration incluse ou les fonctions Lambda pour vous connecter aux services de votre choix.

Intégration avec Service Catalog AppRegistry et Application Manager, une fonctionnalité d'AWS Systems Manager

Cette solution inclut une AppRegistry ressource Service Catalog pour enregistrer le CloudFormation modèle de la solution et ses ressources sous-jacentes en tant qu'application dans AWS Service Catalog AppRegistry et dans AWS Systems Manager Application Manager. Grâce à cette intégration, vous pouvez gérer de manière centralisée les ressources de la solution.

Cas d'utilisation d'Agent Builder vs Bedrock Agent

Cette solution propose deux approches distinctes pour travailler avec des agents d'IA, chacune adaptée à des cas d'utilisation et à des exigences différents :

Fonctionnalité	Cas d'utilisation de Bedrock Agent	Générateur d'agents
Objectif	Invoquez des agents Amazon Bedrock prédéployés	Créez, déployez et gérez des agents personnalisés
Configuration	ID d'agent et ID d'alias uniquement	Configuration complète de l'agent : instructions système, modèles, serveurs MCP, mémoire
Déploiement	Couche d'invocation simple	Cycle de vie complet de l'agent sur AgentCore Runtime
Exécution	Service Amazon Bedrock Agents	SDK Amazon Bedrock AgentCore avec Strands
Intégration d'outils	Configuré dans la console Bedrock Agents	Serveurs MCP (Model Context Protocol) et outils Strands intégrés
Mémoire	Géré par Bedrock Agents (jusqu'à 30 jours)	AgentCore Mémoire avec rétention configurable à court et à long terme
Personnalisation	Limité aux paramètres d'agent prédéployés	Contrôle total des instructions, des modèles, des outils et du comportem ent
Idéal pour	Déploiement rapide des agents existants	Déploiements de production et de développement d'agents personnal isés

> **Note**
>
> Les deux options prennent en charge le streaming en temps réel, l'historique des conversations et une sécurité de niveau professionnel.

Générateur de flux de travail

Workflow Builder permet une orchestration multi-agents en créant un agent superviseur qui délègue le travail à des agents Agent Builder spécialisés. Chaque flux de travail comprend :

- Agent superviseur : L'agent d'entrée qui reçoit les demandes des utilisateurs et coordonne les agents spécialisés
- Agents spécialisés : cas d'utilisation d'Agent Builder auxquels le superviseur peut déléguer des tâches
- Modèle d'agents en tant qu'outils : le superviseur enregistre chaque agent Agent Builder en tant qu'outil et sélectionne de manière autonome les agents à utiliser

Fonctionnalité	Générateur d'agents	Générateur de flux de travail
Objectif	Créez et déployez des agents personnalisés uniques	Orchestrez plusieurs agents Agent Builder
Type d'agent	Agent unique avec outils MCP	Agent superviseur et plusieurs agents Agent Builder
Intégration d'outils	Serveurs MCP et outils Strands	Agents Agent Builder enregistrés en tant qu'outils
Délégation	Invocation directe de l'outil	Sélection et délégation autonomes des agents
Complexité	Tâches à agent unique	Flux de travail multi-agents en plusieurs étapes
Réutilisation des agents	N/A	Réutilise les déploiements d'Agent Builder existants
Idéal pour	Tâches ciblées dans un seul domaine	Des flux de travail complexes nécessitant de multiples spécialis ations

Note

- Les flux de travail nécessitent au moins un cas d'utilisation d'Agent Builder en tant qu'agent agent spécialisé
- Tous les agents spécialisés doivent être des cas d'utilisation d'Agent Builder déployés dans GAAB

Cas d’utilisation

Réponses aux questions sur les données de l'entreprise

LLMs et d'autres modèles de base ont été préentraînés sur un vaste corpus de données, ce qui leur permet de bien exécuter de nombreuses tâches de traitement du langage naturel (NLP). Mais la plupart des modèles de base LLMs sont statiques et ont été préformés, ce qui limite leur capacité à répondre avec précision à des questions sur des sujets nouveaux, spécialisés ou exclusifs. Grâce à l'apprentissage instantané, vous pouvez tirer parti des puissantes fonctionnalités de NLP et de génération de texte d'un LLM pour offrir une expérience client plus riche grâce aux données de votre entreprise.

Prototypage génératif rapide de l'IA

Prête à l'emploi, la solution est fournie avec différents fournisseurs de modèles et cas d'utilisation. Grâce à un assistant de déploiement facile à utiliser, les clients peuvent déployer des cas d'utilisation prédéfinis pour permettre l'expérimentation rapide de différents prototypes et charges de travail d'IA générative.

Comparaison et expérimentation de plusieurs LLM

LLMs fonctionnent différemment, et compte tenu des besoins spécifiques de votre application, vous constaterez peut-être qu'un LLM convient mieux à votre application qu'un autre. Cela peut être dû à des raisons liées à la performance, à la précision, au coût, à la créativité ou à de nombreux autres facteurs. Cette solution vous permet de déployer rapidement plusieurs cas d'utilisation, ce qui vous permet d'expérimenter et de comparer différentes configurations jusqu'à ce que vous trouviez celle qui répond à vos besoins.

Concepts et définitions

Cette section décrit les concepts clés et définit la terminologie spécifique à cette solution :

utilisateur administrateur

Dans le contexte de ce guide, l'utilisateur administrateur est responsable de la gestion du contenu contenu dans le déploiement. Cet utilisateur a accès à l'interface utilisateur du tableau de bord de déploiement et est principalement responsable de l'organisation de l'expérience utilisateur professionnelle. Il s'agit de notre principal client cible.

utilisateur professionnel

Dans le contexte de ce guide, l'utilisateur professionnel représente les personnes pour lesquelles le cas d'utilisation a été déployé. Ils sont les consommateurs de la base de connaissances et le client chargé d'évaluer et d'expérimenter avec la LLMs.

Tableau de bord de déploiement

Le tableau de bord de déploiement est une interface Web qui sert de console de gestion permettant aux utilisateurs administrateurs de visualiser, de gérer et de créer leurs cas d'utilisation. Ce tableau de bord permet aux clients d'expérimenter, d'itérer et de mettre en production rapidement diverses AI/ML charges de travail en tirant parti de ces outils. LLMs

DevOps user

Dans le contexte de ce guide, l' DevOps utilisateur est responsable du déploiement de la solution dans le compte AWS et de la gestion de l'infrastructure, de la mise à jour de la solution, du suivi des performances et du maintien de l'état général et du cycle de vie de la solution.

cas d'utilisation

Les cas d'utilisation sont des applications isolées de la solution globale qui s'intègrent LLMs pour offrir une expérience client plus riche en permettant l'ajout d'une interface en langage naturel aux applications nouvelles ou existantes. Les cas d'utilisation peuvent être déployés via le tableau de bord de déploiement ou de manière autonome.

> **Note**
>
> Pour une référence générale des termes AWS, consultez le glossaire AWS.

Présentation de l’architecture

Cette section fournit des diagrammes d'architecture d'implémentation de référence pour les composants déployés avec cette solution.

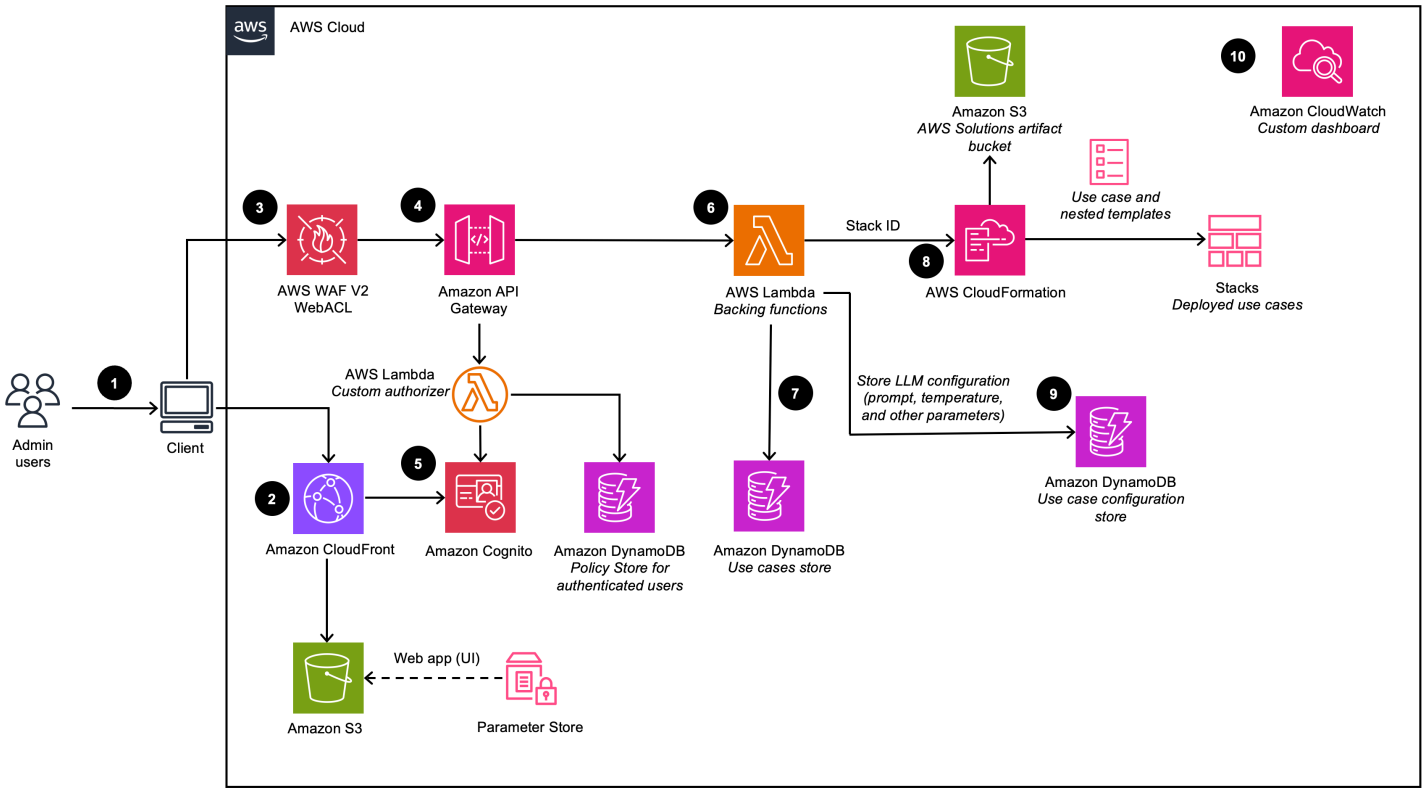
Schémas d'architecture

Pour répondre à de multiples cas d'utilisation et besoins commerciaux, cette solution fournit six CloudFormation modèles AWS :

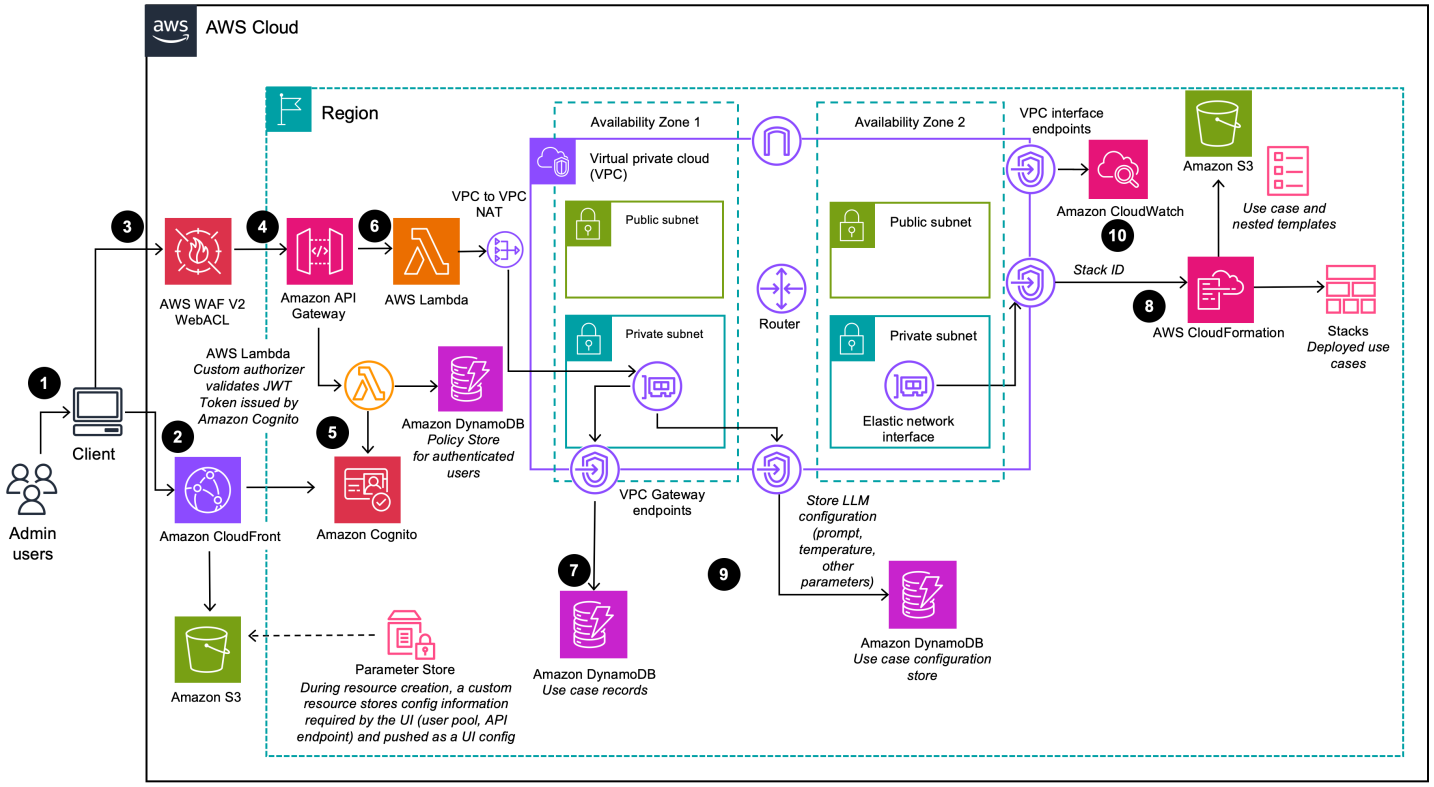
1. Tableau de bord de déploiement - Le tableau de bord de déploiement est une interface Web qui sert de console de gestion permettant aux utilisateurs administrateurs de visualiser, de gérer et de créer leurs cas d'utilisation. Ce tableau de bord permet aux clients d'expérimenter, d'itérer et de mettre en production rapidement diverses AI/ML charges de travail en tirant parti de ces outils. LLMs
2. Cas d'utilisation du texte - Le cas d'utilisation du texte permet aux utilisateurs de découvrir une interface en langage naturel à l'aide de l'IA générative. Ce cas d'utilisation peut être intégré dans des applications nouvelles ou existantes, et peut être déployé via le tableau de bord de déploiement ou indépendamment via une URL fournie.
3. Cas d'utilisation de l'agent Bedrock - Le cas d'utilisation de l'agent Bedrock permet d'utiliser les agents Bedrock existants pour effectuer des tâches ou automatiser des flux de travail répétés.
4. Serveur MCP - Le cas d'utilisation du serveur MCP permet le déploiement et la gestion de serveurs Model Context Protocol qui fournissent un accès standardisé aux outils et aux ressources pour les applications d'IA. Supporte à la fois les méthodes de passerelle pour encapsuler les fonctions Lambda existantes et les serveurs MCP externes APIs, ainsi que les méthodes d'exécution pour déployer des serveurs MCP conteneurisés personnalisés.
5. Agent Builder : l'Agent Builder permet de créer et de déployer des agents d'intelligence artificielle prêts pour la production sur Amazon Bedrock AgentCore avec un contrôle complet de la configuration, l'intégration du serveur MCP et des fonctionnalités de gestion de la mémoire.
6. Générateur de flux de travail : le générateur de flux de travail permet de créer des agents superviseurs qui orchestrent plusieurs agents Agent Builder à l'aide du modèle de délégation Agents as Tools pour les flux de travail multi-agents complexes.

Tableau de bord de déploiement

Décrit l'architecture du tableau de bord de déploiement (en cas de déploiement avec l'option VPC désactivée)



Décrit l'architecture du tableau de bord de déploiement (lorsqu'il est déployé avec l'option VPC activée)



> **Note**
>
> Les CloudFormation ressources AWS sont créées à partir des constructions du kit AWS Cloud Development Kit (AWS CDK).

Le flux de processus de haut niveau pour les composants de solution déployés avec le CloudFormation modèle AWS est le suivant :

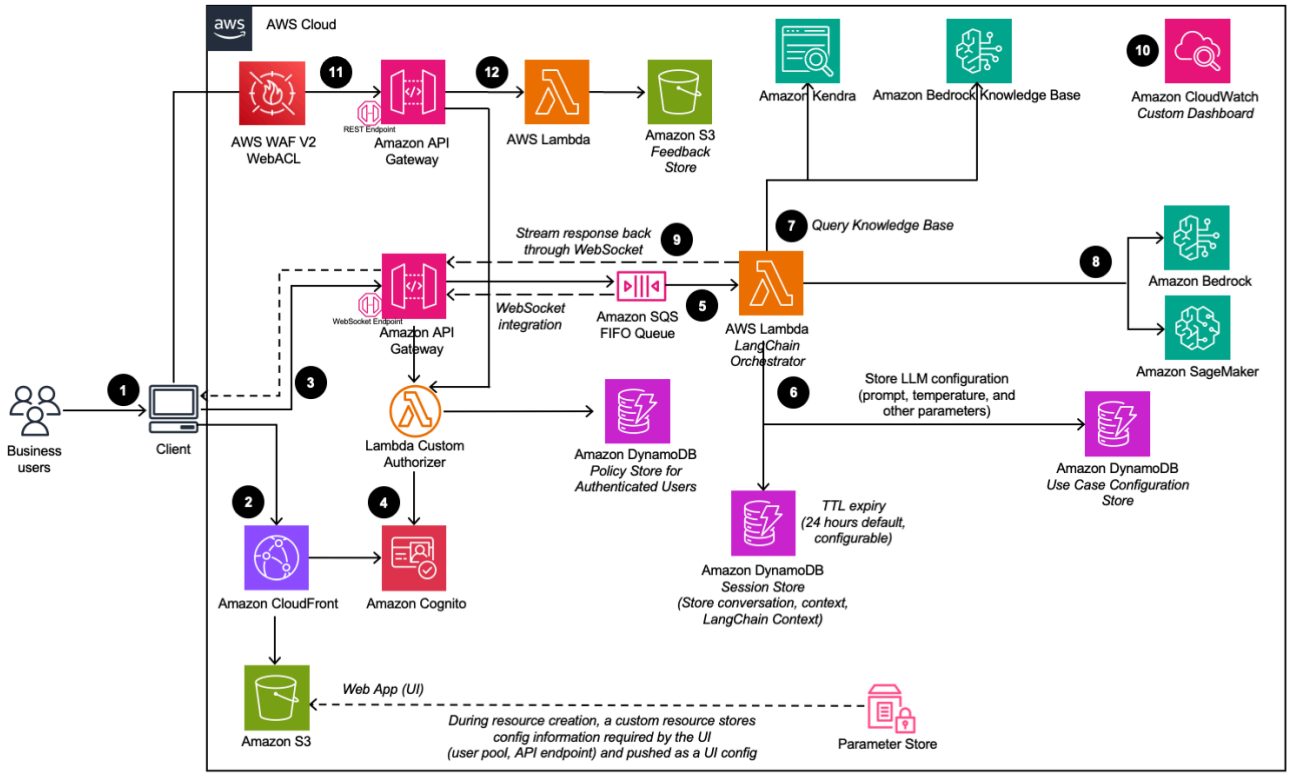
1. Les utilisateurs administrateurs se connectent à l'interface utilisateur (UI) du tableau de bord de déploiement.
2. Amazon CloudFront fournit l'interface utilisateur Web, qui est hébergée dans un compartiment Amazon Simple Storage Service (Amazon S3).
3. AWS WAF les APIs protège des attaques. Cette solution configure un ensemble de règles appelé liste de contrôle d'accès Web (ACL Web) qui autorise, bloque ou compte les requêtes Web en fonction de règles et de conditions de sécurité Web configurables et définies par l'utilisateur.
4. L'interface utilisateur Web utilise un ensemble de REST APIs qui sont exposés à l'aide d'Amazon API Gateway.

5. Amazon Cognito authentifie les utilisateurs et soutient à la fois l'interface utilisateur CloudFront Web et l'API Gateway.
6. AWS Lambda fournit la logique métier pour les points de terminaison REST. Cette fonction Lambda de soutien gère et crée les ressources nécessaires pour effectuer des déploiements de cas d'utilisation à l'aide d'AWS. CloudFormation
7. Amazon DynamoDB stocke la liste des déploiements.
8. Lorsqu'un nouveau cas d'utilisation est créé par l'utilisateur administrateur, la fonction Lambda de soutien lance un événement de création de CloudFormation pile pour le cas d'utilisation demandé.
9. Toutes les options de configuration LLM fournies par l'utilisateur administrateur dans l'assistant de déploiement sont enregistrées dans DynamoDB. Le déploiement utilise cette table DynamoDB pour configurer le LLM lors de l'exécution.
10. À l'aide d'Amazon CloudWatch, cette solution collecte des métriques opérationnelles auprès de divers services afin de générer des tableaux de bord personnalisés qui vous permettent de surveiller les performances et la santé opérationnelle de la solution.

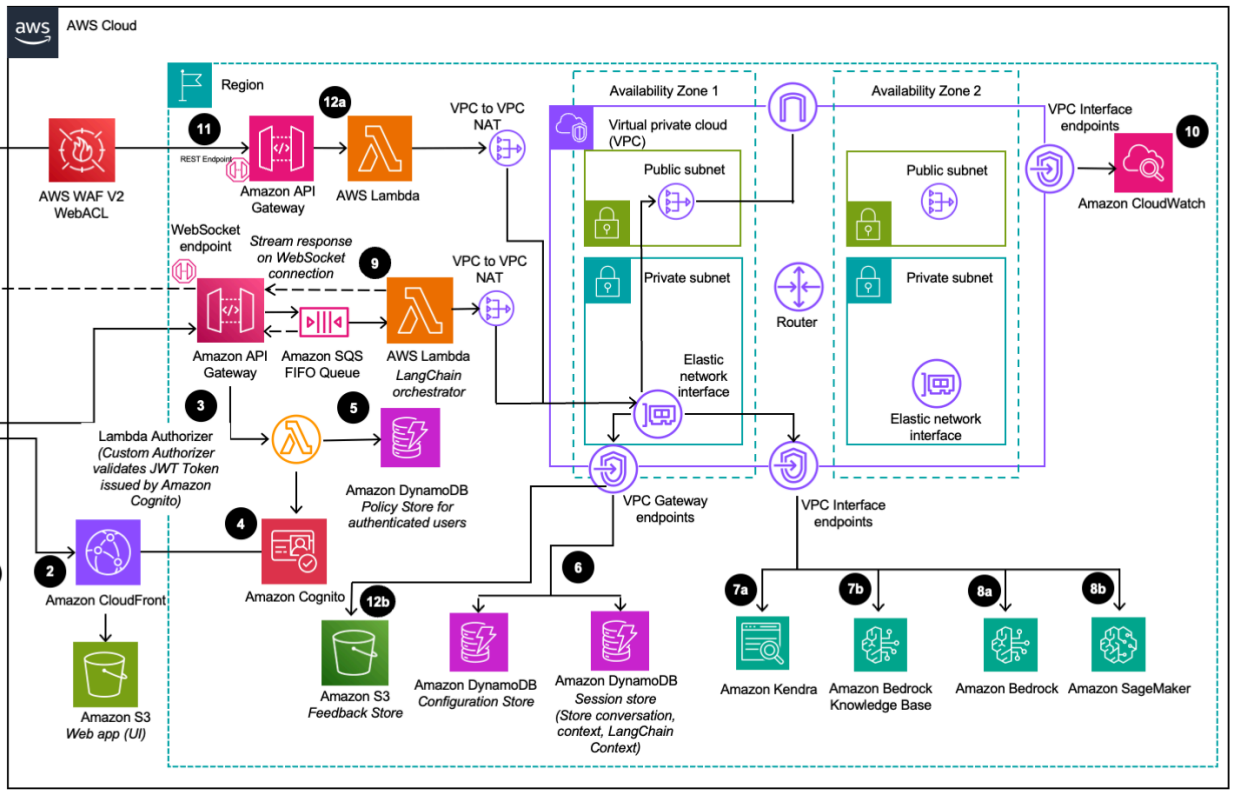
> **Note**
>
> - Si vous choisissez de déployer cette solution dans un Amazon VPC, les données seront acheminées au sein de votre réseau privé.
> - Bien que le tableau de bord de déploiement puisse être lancé dans la plupart des régions AWS, les cas d'utilisation déployés sont soumis à certaines restrictions en fonction de la disponibilité des services. Consultez la section Régions AWS prises en charge pour plus de détails.

Cas d'utilisation du texte

Décrit l'architecture des cas d'utilisation du texte (en cas de déploiement avec l'option VPC désactivée)



Décrit l'architecture des cas d'utilisation du texte (lorsqu'elle est déployée avec l'option VPC activée)



Le flux de processus de haut niveau pour les composants de solution déployés avec le CloudFormation modèle AWS est le suivant :

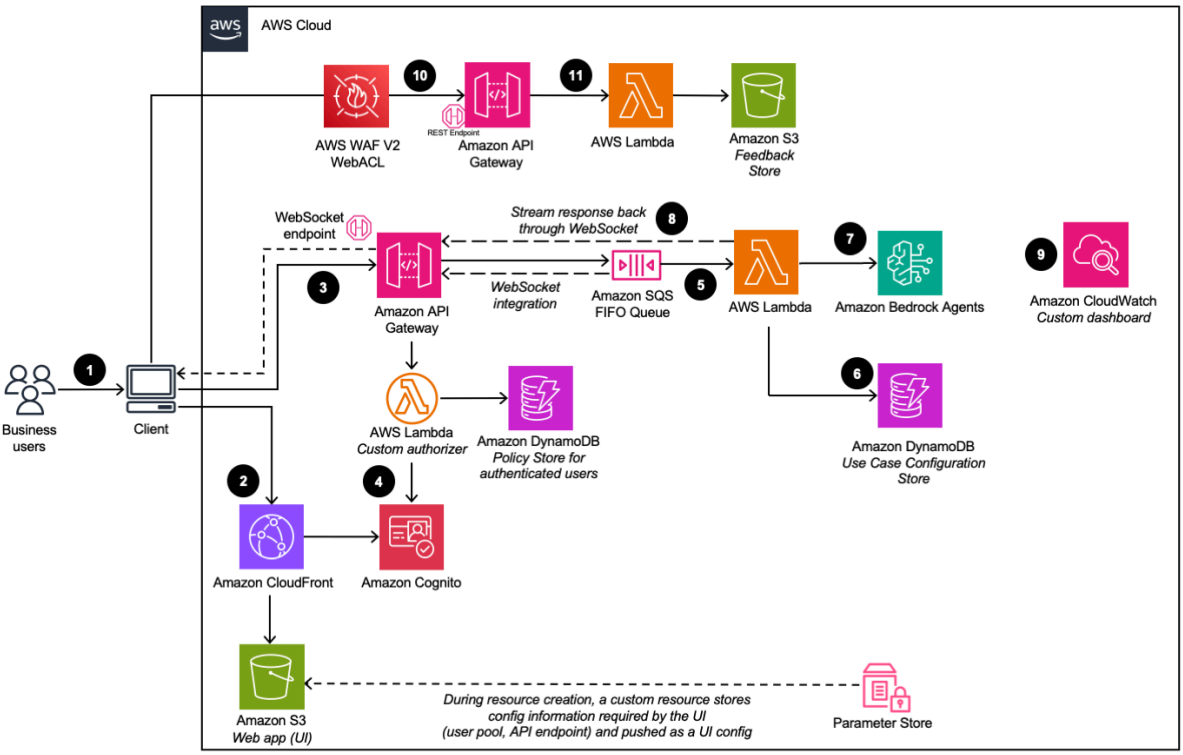
1. Les utilisateurs administrateurs déploient le cas d'utilisation à l'aide du tableau de bord de déploiement. Les utilisateurs professionnels se connectent à l'interface utilisateur du cas d'utilisation.
2. CloudFront fournit l'interface utilisateur Web hébergée dans un compartiment S3.
3. L'interface utilisateur Web tire parti d'une WebSocket intégration créée à l'aide d'API Gateway. L'API Gateway est soutenu par une fonction d'autorisation Lambda personnalisée, qui renvoie la politique AWS Identity and Access Management (IAM) appropriée en fonction du groupe Amazon Cognito auquel appartient l'utilisateur authentifié. La politique est stockée dans DynamoDB.
4. Amazon Cognito authentifie les utilisateurs et soutient à la fois l'interface utilisateur CloudFront Web et l'API Gateway.
5. Les demandes entrantes de l'utilisateur professionnel sont transmises d'API Gateway à une file d'attente Amazon SQS, puis à l'LangChain orchestrateur. L'LangChain orchestrateur est un ensemble de fonctions et de couches Lambda qui fournissent la logique métier permettant de répondre aux demandes émanant de l'utilisateur professionnel. La file d'attente permet le fonctionnement asynchrone de l'intégration entre API Gateway et Lambda. La file d'attente transmet les informations de connexion aux fonctions Lambda qui publieront ensuite les résultats directement sur la connexion Websocket d'API Gateway afin de prendre en charge les appels d'inférence de longue durée.
6. L'LangChain orchestrateur utilise Amazon DynamoDB pour obtenir les options LLM configurées et les informations de session nécessaires (telles que l'historique des discussions).
7. Si une base de connaissances est activée pour le déploiement, l'LangChain orchestrateur utilise Amazon Kendra ou Knowledge Bases for Amazon Bedrock pour exécuter une requête de recherche afin de récupérer des extraits de documents.
8. À l'aide de l'historique des discussions, de la requête et du contexte de la base de connaissances, l'LangChain orchestrateur crée l'invite finale et envoie la demande au LLM hébergé sur Amazon Bedrock ou Amazon AI. SageMaker
9. Lorsque la réponse provient du LLM, l'LangChain orchestrateur la renvoie via l'API Gateway WebSocket pour être consommée par l'application cliente.
10. À l'aide d'Amazon CloudWatch, cette solution collecte des métriques opérationnelles auprès de différents services afin de générer des tableaux de bord personnalisés qui vous permettent de surveiller les performances et la santé opérationnelle du déploiement.

11. Si la collecte de commentaires est activée, un point de terminaison d'API REST, exploitant Amazon API Gateway, est mis à disposition pour recueillir les commentaires des utilisateurs.
12. Le feedback, qui soutient Lambda, complète le feedback envoyé avec des métadonnées supplémentaires spécifiques au cas d'utilisation (par exemple, le modèle utilisé) et stocke les données dans Amazon S3 pour une analyse et des rapports ultérieurs par les DevOps utilisateurs.

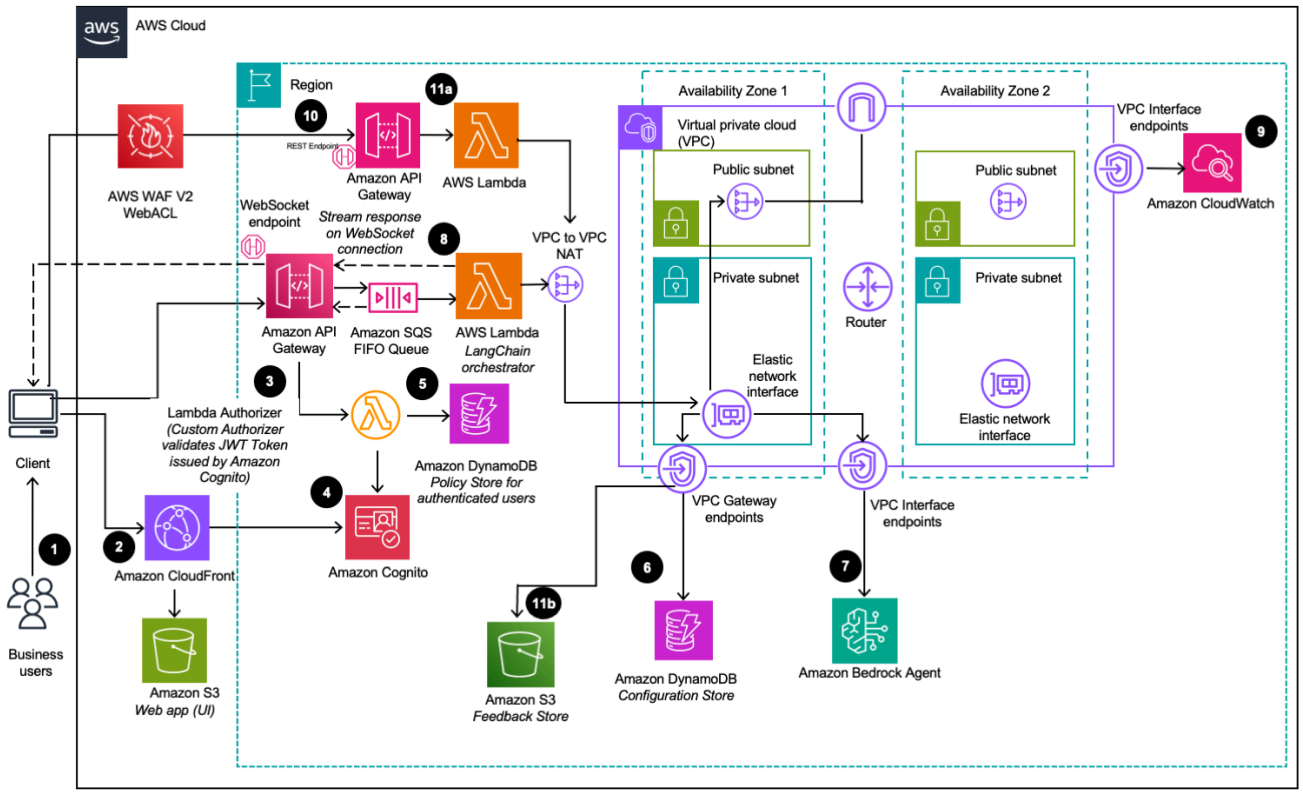
> **Note**
>
> Si vous choisissez de déployer cette solution dans un Amazon VPC, les données seront acheminées vers votre réseau privé.

Cas d'utilisation de Bedrock Agent

Décrit l'architecture du cas d'utilisation de Bedrock Agent (lorsqu'il est déployé avec l'option VPC désactivée)



Décrit l'architecture du cas d'utilisation de Bedrock Agent (lorsqu'il est déployé avec l'option VPC activée)



Le flux de processus de haut niveau pour les composants de solution déployés avec le CloudFormation modèle AWS est le suivant :

1. Les utilisateurs administrateurs déploient le cas d'utilisation à l'aide du tableau de bord de déploiement. Les utilisateurs professionnels se connectent à l'interface utilisateur du cas d'utilisation.
2. CloudFront fournit l'interface utilisateur Web hébergée dans un compartiment S3.
3. L'interface utilisateur Web tire parti d'une WebSocket intégration créée à l'aide d'API Gateway. L'API Gateway est soutenu par une fonction d'autorisation Lambda personnalisée, qui renvoie la politique AWS Identity and Access Management (IAM) appropriée en fonction du groupe Amazon Cognito auquel appartient l'utilisateur authentifié. La politique est stockée dans DynamoDB.
4. Amazon Cognito authentifie les utilisateurs et soutient à la fois l'interface utilisateur CloudFront Web et l'API Gateway.
5. Les demandes entrantes de l'utilisateur professionnel sont transmises d'API Gateway à une file d'attente Amazon SQS, puis à la fonction AWS Lambda. La file d'attente permet le fonctionnement asynchrone de l'intégration entre API Gateway et Lambda. La file d'attente transmet les informations de connexion à la fonction Lambda qui publiera ensuite les résultats directement

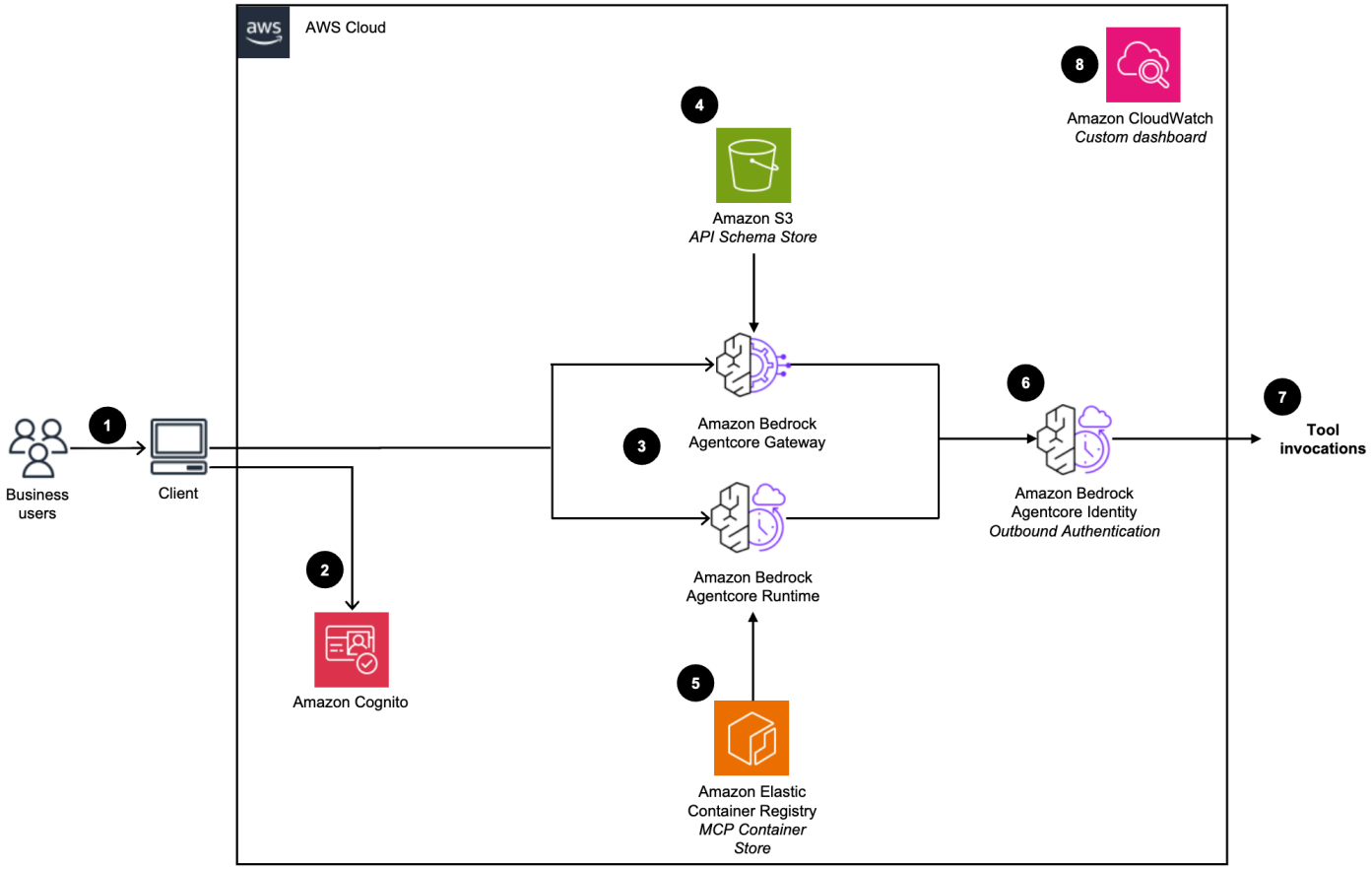
sur la connexion Websocket d'API Gateway afin de prendre en charge les appels d'inférence de longue durée.

6. La fonction AWS Lambda utilise Amazon DynamoDB pour obtenir les configurations de cas d'utilisation nécessaires
7. À l'aide de l'entrée utilisateur et de toute configuration de cas d'utilisation pertinente, la fonction AWS Lambda crée et envoie une charge utile de demande à l'agent Amazon Bedrock configuré afin de répondre à l'intention de l'utilisateur.
8. Lorsque la réponse provient de l'agent Amazon Bedrock, la fonction Lambda la renvoie via l'API WebSocket Gateway pour être consommée par l'application cliente.
9. À l'aide d'Amazon CloudWatch, cette solution collecte des métriques opérationnelles auprès de différents services afin de générer des tableaux de bord personnalisés qui vous permettent de surveiller les performances et la santé opérationnelle du déploiement.
10. Si la collecte de commentaires est activée, un point de terminaison d'API REST, exploitant Amazon API Gateway, est mis à disposition pour recueillir les commentaires des utilisateurs.
11. Le feedback, qui soutient Lambda, complète le feedback envoyé avec des métadonnées supplémentaires spécifiques aux cas d'utilisation et stocke les données dans Amazon S3 pour une analyse et des rapports ultérieurs par les DevOps utilisateurs.

> **Note**
>
> Si vous choisissez de déployer cette solution dans un Amazon VPC, les données seront acheminées au sein de votre réseau privé.

Cas d'utilisation du serveur MCP

Décrit l'architecture des cas d'utilisation du serveur MCP



Le cas d'utilisation du serveur MCP permet le déploiement et la gestion de serveurs Model Context Protocol sur Amazon AgentCore Bedrock. Les serveurs MCP fournissent une interface standardisée permettant aux applications d'IA d'accéder aux outils, aux ressources et aux sources de données d'entreprise.

La solution prend en charge deux méthodes de déploiement :

- Méthode de passerelle : transforme les fonctions Lambda existantes, les serveurs APIs REST ou les serveurs MCP externes en outils MCP, gérant automatiquement la traduction des protocoles
- Méthode d'exécution : déploie des serveurs MCP conteneurisés personnalisés à partir d'images Amazon ECR

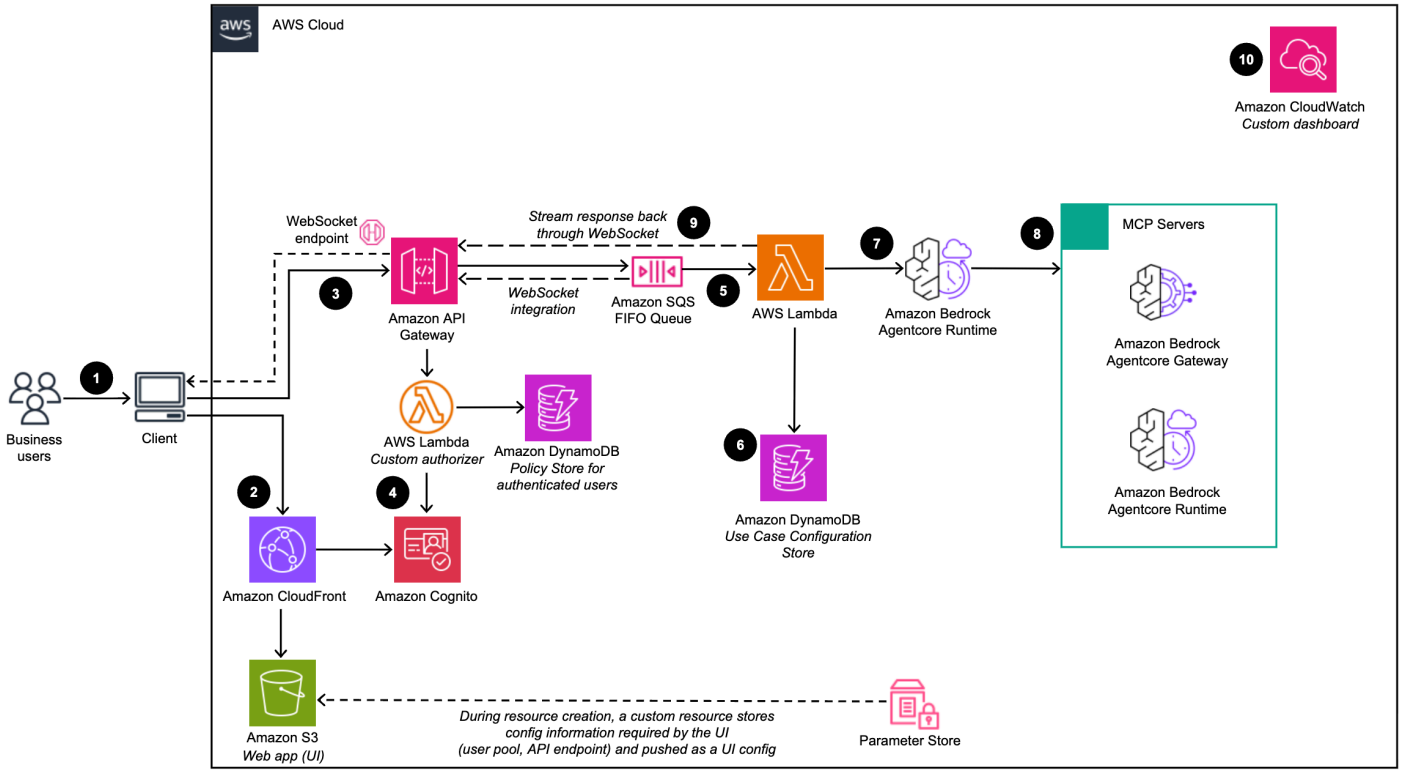
Le flux de processus de haut niveau pour le déploiement du serveur MCP est le suivant :

1. Les utilisateurs administrateurs déploient le cas d'utilisation du serveur MCP à l'aide du tableau de bord de déploiement, en sélectionnant la méthode de déploiement Gateway ou Runtime.

2. Cette action est authentifiée avec Amazon Cognito.
3. Pour le déploiement de la passerelle, la solution crée une AgentCore passerelle Amazon Bedrock qui transforme les fonctions APIs Lambda existantes ou les serveurs MCP externes en outils compatibles avec le protocole MCP. Pour le déploiement du Runtime, la solution déploie des serveurs MCP conteneurisés sur Amazon Bedrock AgentCore Runtime à l'aide des images ECR fournies.
4. Les déploiements de passerelles récupèrent les API/Lambda/Smithy schémas nécessaires depuis leur emplacement de téléchargement dans Amazon S3 ou se connectent directement aux points de terminaison URL du serveur MCP.
5. Les déploiements d'exécution extraient le serveur MCP conteneurisé fourni par l'utilisateur depuis Amazon Elastic Container Registry (ECR)
6. Le serveur MCP est équipé d'un client Amazon Bedrock Identity AgentCore OAuth
7. Le serveur MCP met les outils associés à disposition sur le point de terminaison /mcp pour que les agents puissent les découvrir.
8. Amazon CloudWatch collecte des statistiques opérationnelles et des journaux à partir des déploiements de serveurs MCP à des fins de surveillance et de résolution des problèmes.

Cas d'utilisation d'Agent Builder

Décrit l'architecture d'Agent Builder



Le flux de processus de haut niveau pour les composants Agent Builder déployés avec le CloudFormation modèle AWS est le suivant :

1. Les utilisateurs administrateurs déploient le cas d'utilisation à l'aide du tableau de bord de déploiement. Les utilisateurs professionnels se connectent à l'interface utilisateur du cas d'utilisation.
2. CloudFront fournit l'interface utilisateur Web hébergée dans un compartiment S3.
3. L'interface utilisateur Web tire parti d'une WebSocket intégration créée à l'aide d'API Gateway. L'API Gateway est soutenu par une fonction d'autorisation Lambda personnalisée, qui renvoie la politique AWS Identity and Access Management (IAM) appropriée en fonction du groupe Amazon Cognito auquel appartient l'utilisateur authentifié. La politique est stockée dans DynamoDB.
4. Amazon Cognito authentifie les utilisateurs et soutient à la fois l'interface utilisateur CloudFront Web et l'API Gateway.
5. Les demandes entrantes de l'utilisateur professionnel sont transmises d'API Gateway à une file d'attente Amazon SQS, puis à la fonction AWS Lambda. La file d'attente permet le fonctionnement asynchrone de l'intégration entre API Gateway et Lambda. La file d'attente transmet les informations de connexion à la fonction Lambda qui publiera ensuite les résultats directement

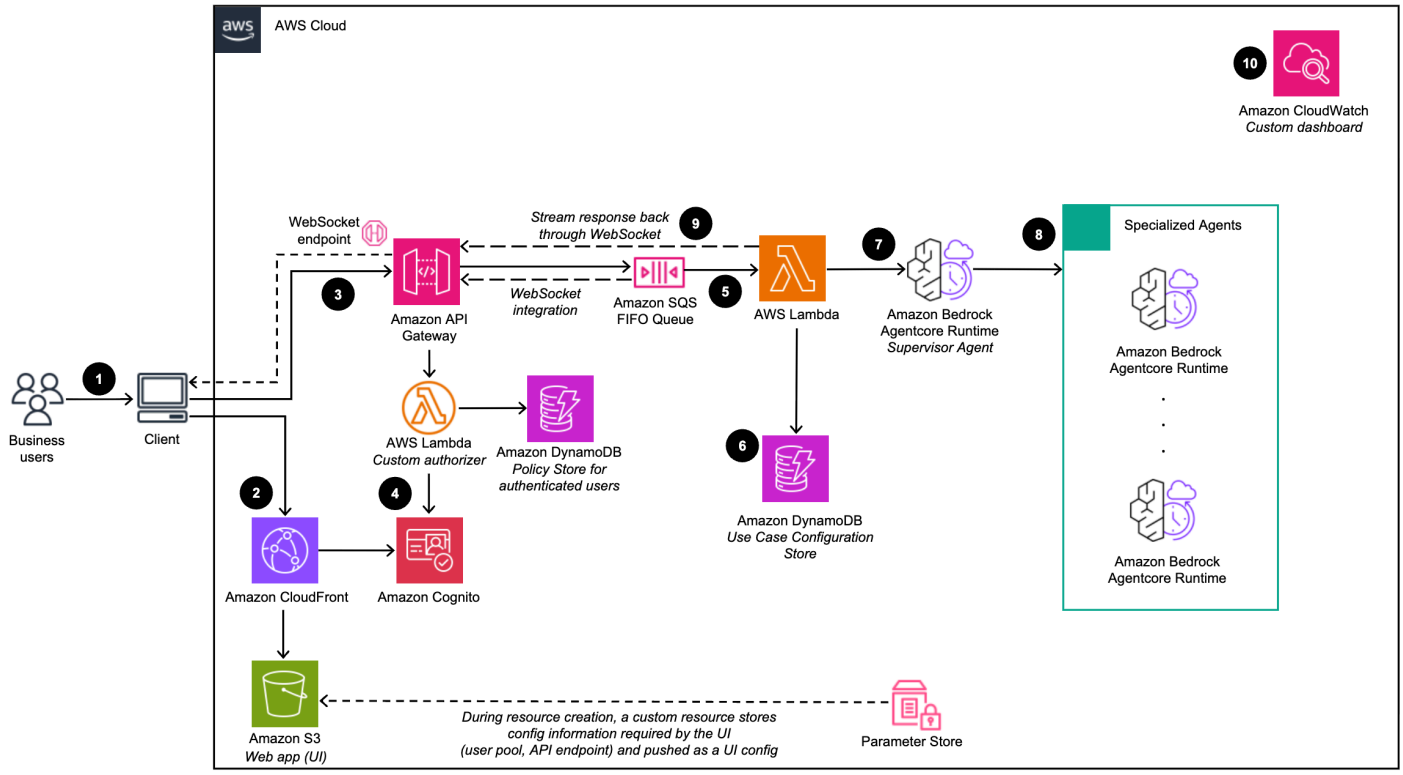
sur la connexion Websocket d'API Gateway afin de prendre en charge les appels d'inférence de longue durée.

6. La fonction AWS Lambda extrait la configuration de l'agent depuis DynamoDB.
7. À l'aide de l'entrée utilisateur et de toute configuration de cas d'utilisation pertinente, la fonction AWS Lambda crée et envoie une charge utile de demande à l'agent, qui s'exécute sur Amazon Bedrock Runtime. AgentCore
8. L'agent se connecte aux serveurs MCP associés et enregistre les outils sur l'instance de l'agent Strands. L'agent sélectionne et exécute ensuite des actions de manière autonome en fonction des descriptions des outils et des exigences des tâches.
9. Lorsque la réponse provient de l' AgentCore environnement d'exécution Amazon Bedrock, la fonction Lambda la renvoie via l'API WebSocket Gateway pour être consommée par l'application cliente.

> **Note**
>
> - Le traitement de l'agent est limité au délai d'exécution de Lambda (15 minutes).

Cas d'utilisation de Workflow Builder

Décrit l'architecture de Workflow Builder



Le flux de processus de haut niveau pour les composants de Workflow Builder déployés avec le CloudFormation modèle AWS est le suivant :

1. Les utilisateurs administrateurs déploient le flux de travail à l'aide du tableau de bord de déploiement, en sélectionnant les agents Agent Builder à inclure en tant qu'agents spécialisés.
2. CloudFront fournit l'interface utilisateur Web hébergée dans un compartiment S3.
3. L'interface utilisateur Web tire parti d'une WebSocket intégration créée à l'aide d'API Gateway. L'API Gateway est soutenu par une fonction d'autorisation Lambda personnalisée, qui renvoie la politique AWS Identity and Access Management (IAM) appropriée en fonction du groupe Amazon Cognito auquel appartient l'utilisateur authentifié. La politique est stockée dans DynamoDB.
4. Amazon Cognito authentifie les utilisateurs et soutient à la fois l'interface utilisateur CloudFront Web et l'API Gateway.
5. Les demandes entrantes de l'utilisateur professionnel sont transmises d'API Gateway à une file d'attente Amazon SQS, puis à la fonction AWS Lambda. La file d'attente permet le fonctionnement asynchrone de l'intégration entre API Gateway et Lambda.
6. La fonction AWS Lambda extrait la configuration du flux de travail depuis DynamoDB, y compris la liste des agents Agent Builder spécialisés.

7. À l'aide de la saisie utilisateur et de la configuration du flux de travail, Lambda envoie des demandes à l'Amazon Bedrock AgentCore Runtime hébergeant l'agent de supervision.
8. L'agent superviseur crée des instances locales de tous les agents Agent Builder spécialisés dans l'environnement AgentCore d'exécution. Ces agents spécialisés sont enregistrés en tant qu'outils en utilisant le modèle Agents as Tools. Le superviseur sélectionne et délègue ensuite le travail de manière autonome à des agents spécialisés en fonction des descriptions des agents et des exigences des tâches.
9. L'agent superviseur agrège les résultats des agents spécialisés et formule la réponse finale, la renvoyant au Lambda pour être retransmise à l'application cliente via le Websocket API Gateway.

> **Note**
>
> - Le traitement du flux de travail est limité au délai d'exécution de Lambda (15 minutes).

Considérations relatives à la conception d'AWS Well-Architected

Cette solution a été conçue selon les meilleures pratiques de l'AWS Well-Architected Framework, qui aide les clients à concevoir et à exploiter des charges de travail fiables, sécurisées, efficaces et rentables dans le cloud.

Cette section décrit comment les principes de conception et les meilleures pratiques du Well-Architected Framework ont été appliqués lors de la création de cette solution.

Excellence opérationnelle

Cette section décrit comment nous avons conçu cette solution en utilisant les principes et les meilleures pratiques du pilier de l'excellence opérationnelle.

- Nous avons conçu la solution en infrastructure-as-code utilisant Amazon CloudFormation.
- Les fonctions Lambda transmettent des métriques personnalisées CloudWatch et un tableau de CloudWatch bord personnalisé pour surveiller l'état de santé de la solution.
- Les composants de la solution sont hautement modularisés, ce qui permet de choisir les composants à déployer.

Sécurité

Cette section décrit comment nous avons conçu cette solution en utilisant les principes et les meilleures pratiques du pilier de sécurité.

- Le tableau de bord de déploiement et tous les cas d'utilisation sont authentifiés et autorisés avec Amazon Cognito.
- Toutes les communications interservices utilisent des rôles AWS IAM.
- Tous les rôles de solution suivent le principe du moindre privilège, ce qui signifie que seules les autorisations minimales requises sont accordées.
- Tout le stockage de données, y compris les compartiments S3, DynamoDB et Amazon Kendra, est crypté au repos.

Fiabilité

Cette section décrit comment nous avons conçu cette solution en utilisant les principes et les meilleures pratiques du pilier de fiabilité.

- Architecture basée sur le paradigme sans serveur.
- Nous avons conçu l'architecture pour une évolutivité horizontale à la demande et une reprise automatique en cas de défaillance de l'infrastructure sous-jacente.
- L'architecture inclut la mise en mémoire tampon et la limitation des demandes afin de ne pas surcharger les points de terminaison sous-jacents.

Efficacité des performances

Cette section décrit comment nous avons conçu cette solution en utilisant les principes et les meilleures pratiques du pilier de l'efficacité des performances.

- La solution utilise DynamoDB, une base de données NoSQL sans serveur entièrement gérée avec un dimensionnement à la demande.
- La solution utilise Amazon S3 pour le stockage d'objets et pour héberger un site Web (via CloudFront) afin de fournir un faible coût, une évolutivité et une durabilité de 11 à 9 secondes.

Optimisation des coûts

Cette section décrit comment nous avons conçu cette solution en utilisant les principes et les meilleures pratiques du pilier d'optimisation des coûts.

- Dans la mesure du possible, nous avons conçu la solution pour utiliser une architecture sans serveur ; vous ne payez donc que pour ce que vous utilisez.

Durabilité

Cette section décrit comment nous avons conçu cette solution en utilisant les principes et les meilleures pratiques du pilier du développement durable.

- L'architecture modulaire et composée de composants de la solution offre la flexibilité nécessaire pour personnaliser les ressources à allouer pour des cas d'utilisation individuels.
- L'architecture utilise le calcul et le stockage sans serveur, ce qui optimise l'utilisation des ressources.
- En tant que solution basée sur le cloud, cette solution bénéficie du partage des ressources, du réseau, du refroidissement électrique et des installations physiques.

Détails de l'architecture

Cette section décrit les composants et les services AWS qui constituent cette solution ainsi que les détails de l'architecture sur la manière dont ces composants fonctionnent ensemble.

Services AWS inclus dans cette solution

Service AWS	Description
Amazon API Gateway	Noyau. Ce service fournit le REST APIs pour le tableau de bord de déploiement et l' WebSocket API pour le cas d'utilisation.
AWS CloudFormation	Noyau. Cette solution est distribuée sous forme de CloudFormation modèle et CloudFormation déploie les ressources AWS associées à la solution.
Amazon CloudFront	Noyau. CloudFront diffuse le contenu Web hébergé dans Amazon S3.
Amazon Cognito	Noyau. Ce service gère la gestion des utilisate urs et l'authentification pour l'API.
Amazon DynamoDB	Noyau. DynamoDB stocke les informations de déploiement et les détails de configuration pour le tableau de bord de déploiement. Il stocke l'historique des discussions et des conversat ions IDs dans le cas d'utilisation du texte pour permettre l'historique des conversations et la désambiguïsation des requêtes.
AWS Lambda	Noyau. La solution utilise les fonctions Lambda pour : * Soutenir les points de terminaison WebSocket REST et API * Gérer la logique de base de

Service AWS	Description
	chaque orchestrateur de cas d'utilisation * Implémenter des ressources personnalisées lors du déploiement CloudFormation
Amazon S3	Noyau. Amazon S3 héberge le contenu Web statique.
Amazon CloudWatch	Soutenir. Cette solution publie les journaux des ressources de la solution dans les CloudWatch journaux, et publie les métriques dans les CloudWatch métriques. La solution crée également un CloudWatch tableau de bord pour consulter ces données.
AWS Systems Manager	Soutenir. Systems Manager assure la surveillance des ressources au niveau de l'application et la visualisation des opérations sur les ressources et des données de coûts. Également utilisé pour stocker les données de configuration dans le Parameter Store.
AWS WAF	Soutenir. AWS WAF est déployé devant le déploiement d'API Gateway pour le protéger.
Amazon Bedrock	Facultatif. La solution s'appuie sur Amazon Bedrock pour accéder à des modèles de base ou personnalisés, à Amazon Bedrock Agents et à des bases de connaissances Amazon Bedrock. Amazon Bedrock est l'intégration recommandée pour empêcher vos données de quitter le réseau AWS.
Amazon Bedrock AgentCore	Facultatif La solution utilise Amazon Bedrock AgentCore pour exécuter et prendre en charge les connexions au serveur MCP ainsi que les cas d'utilisation d'Agent Builder et de Workflow.

Service AWS	Description
Amazon Elastic Container Registry (Amazon ECR)	Facultatif. Pour les déploiements d'Agent Builder, ECR stocke et distribue des images de conteneurs d'agents. La solution utilise le cache ECR Pull-Through pour récupérer automatiq uement des images d'agents prédéfinies à partir du référentiel ECR public de l'équipe GAAB.
AWS Distro pour OpenTelemetry (ADOT)	Facultatif. Pour les déploiements d'Agent Builder, ADOT fournit une instrumentation automatique pour l'observabilité des agents, permettant un suivi distribué et une journalis ation structurée des opérations des agents.
Amazon Kendra	Facultatif. Dans le cas d'utilisation du texte, les utilisateurs administrateurs peuvent éventuell ement décider de connecter un index Amazon Kendra à utiliser comme base de connaissa nces pour la conversation avec le LLM. Cela peut être utilisé pour injecter de nouvelles informations dans le LLM, lui donnant la possibilité d'utiliser ces informations dans ses réponses.

Service AWS	Description
Amazon SageMaker AI	Facultatif. La solution peut s'intégrer à un point de terminaison d'inférence Amazon SageMaker AI pour accéder FMs à des accès hébergés dans votre compte et votre région AWS. Il s'agit d'une intégration privilégiée pour empêcher vos données de quitter le réseau AWS. **Note** Vous devez déployer la solution dans la même région où le point de terminaison d'inférence est disponible.
Amazon Virtual Private Cloud	Facultatif. La solution offre la possibilité de déployer des composants avec une configuration compatible VPC. Lorsque vous déployez la solution avec une configuration compatible VPC, vous avez la possibilité de laisser la solution créer un VPC pour vous ou d'utiliser un VPC existant qui existe dans le même compte et dans la même région que ceux où la solution sera déployée (Bring Your Own VPC). Si la solution crée le VPC, elle crée les composants réseau nécessaires, notamment les sous-réseaux, les groupes de sécurité et leurs règles, les tables de routage, le réseau, les passerelles NAT ACLs, les passerelles Internet, les points de terminaison VPC et ses politiques.

Tableau de bord de déploiement

Autorisateurs personnalisés API Gateway

En apparence, les autorisateurs Lambda personnalisés pour API Gateway sont utilisés pour tous les appels d'API (WebSocket basés ou RESTful non) afin de valider si un utilisateur donné est autorisé à effectuer une action en fonction du ou des groupes auxquels il appartient. Cet autorisateur personnalisé est soutenu par une table DynamoDB contenant les politiques de chaque groupe. Lors de l'appel d'une API, API Gateway invoque la fonction Lambda d'autorisation personnalisée, qui décode le jeton d'accès Amazon Cognito fourni afin de déterminer à quels groupes d'utilisateurs appartient l'utilisateur. La table des politiques est ensuite interrogée par nom de groupe afin de renvoyer la politique appropriée pour ce groupe.

À chaque nouveau déploiement de cas d'utilisation, la politique d'administration est mise à jour pour stocker une nouvelle instruction autorisant l'action Execute-API:Invoke sur l'API de ce cas d'utilisation. Lorsque des cas d'utilisation sont supprimés, l'instruction correspondante est supprimée de la politique.

Pour les groupes créés pour un cas d'utilisation individuel, une seule instruction est présente dans la politique, autorisant l'action Execute-API:Invoke uniquement sur l'API de ce cas d'utilisation.

Grâce à cette structure, tout utilisateur appartenant au groupe d'un cas d'utilisation peut accéder à l'API de ce cas d'utilisation. Un seul utilisateur peut également être ajouté manuellement à plusieurs groupes pour lui permettre d'utiliser plusieurs cas d'utilisation.

> **Warning**
>
> Vous pouvez également modifier manuellement les politiques d'un groupe donné dans le tableau des politiques si vous souhaitez accorder l'accès à un nouveau cas d'utilisation à un groupe d'utilisateurs existant. Le groupe de cas d'utilisation est supprimé lorsque le cas d'utilisation est supprimé (même si vous avez effectué des modifications manuelles). Soyez donc prudent lorsque vous supprimez un cas d'utilisation.

Dans le cas où une pile de cas d'utilisation est déployée de manière autonome (sans utiliser le tableau de bord de déploiement), un groupe d'utilisateurs Amazon Cognito est créé pour ce déploiement, contenant un seul utilisateur ayant accès à l'API de ce cas d'utilisation. Ce groupe d'utilisateurs appartient uniquement à ce cas d'utilisation et n'est pas partagé entre d'autres déploiements autonomes.

Cas d'utilisation du texte

Support de streaming

Dans une application de chat, la latence est un indicateur important pour garantir une expérience utilisateur réactive. La possibilité que les inférences du LLM prennent de quelques secondes à quelques minutes pose des défis quant à la meilleure façon de proposer du contenu aux clients. Pour cette raison, plusieurs fournisseurs de LLM autorisent le renvoi des réponses en continu à l'appelant. Au lieu d'attendre que l'inférence soit complète avant de renvoyer une réponse, chaque jeton peut être renvoyé lorsqu'il est disponible.

Pour faciliter l'utilisation de cette fonctionnalité, l'exemple d'utilisation du texte a été conçu pour utiliser une WebSocket API afin de renforcer l'expérience de chat. Ceci WebSocket est déployé via API Gateway. L'utilisation d'une WebSocket API permet de créer une connexion au début d'une session de chat et de diffuser les réponses via ce socket. Cela permet aux applications frontales d'offrir une meilleure expérience utilisateur.

> **Note**
>
> Même si un modèle prend en charge le streaming, cela ne signifie pas nécessairement que la solution sera en mesure de renvoyer les réponses via l' WebSocket API. La solution doit activer une logique personnalisée pour prendre en charge le streaming pour chaque fournisseur de modèles. Si le streaming est disponible, les utilisateurs administrateurs pourront accéder à enable/disable cette fonctionnalité au moment du déploiement.

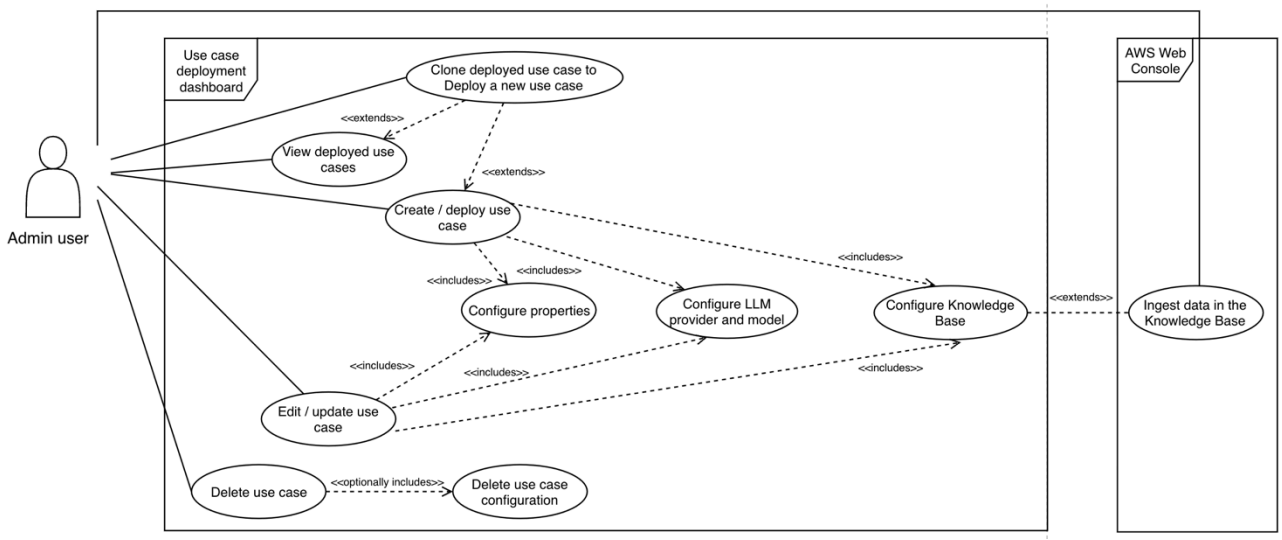
Fonctionnement de la solution Generative AI Application Builder sur AWS

L'utilisateur administrateur s'interface principalement avec le tableau de bord de déploiement pour visualiser, créer et gérer des déploiements de cas d'utilisation nouveaux et existants. Grâce à ce tableau de bord, l'utilisateur administrateur a accès aux actions suivantes :

- Afficher la liste des déploiements
- Créez de nouveaux déploiements
- Modifier les déploiements existants
- Cloner la configuration d'un déploiement pour créer un nouveau déploiement

- Supprimer un déploiement (déprovisionner les ressources par le biais d'une CloudFormation suppression)
- Supprimer définitivement les détails de configuration d'un déploiement

Représente le schéma de cas d'utilisation pour l'utilisateur administrateur du tableau de bord de déploiement



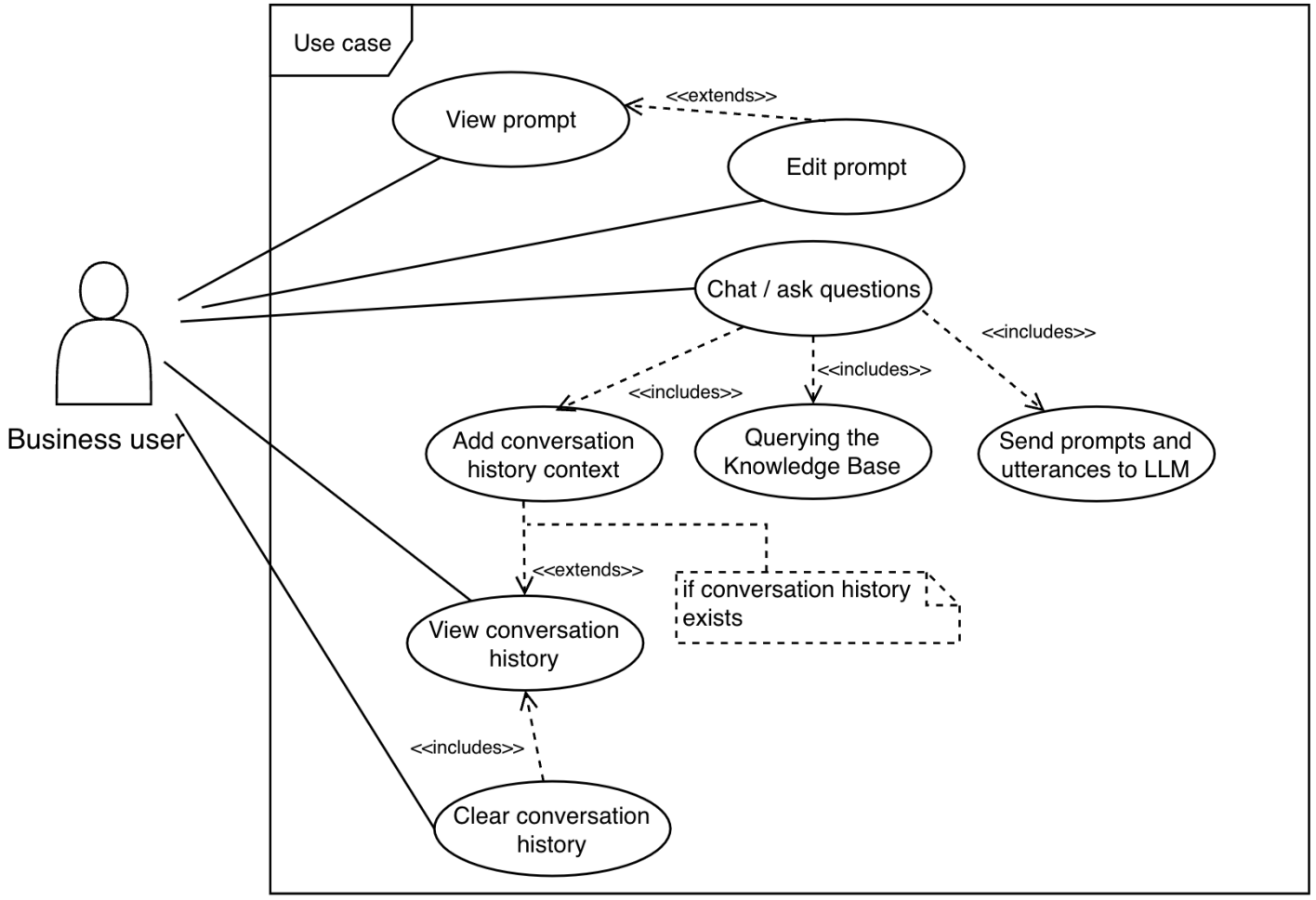
> **Note**
>
> L'utilisateur administrateur ne dispose peut-être pas d'un accès direct à la console AWS. Dans ce cas, l'utilisateur administrateur doit travailler avec l' DevOps utilisateur pour prendre en charge des actions telles que l'ingestion de données dans une base de connaissances Kendra.

Pour le cas d'utilisation du texte, l'utilisateur professionnel a accès à une interface utilisateur lui permettant de discuter avec le LLM. Les spécificités de cette configuration sont contrôlées par les paramètres de déploiement configurés par l'utilisateur administrateur. Dans le cas d'utilisation du texte, l'utilisateur professionnel a accès aux actions suivantes :

- Envoyer des messages via l'interface de chat
- Afficher l'historique des conversations
- Effacer l'historique des conversations
- Afficher l'invite

- Invite de modification

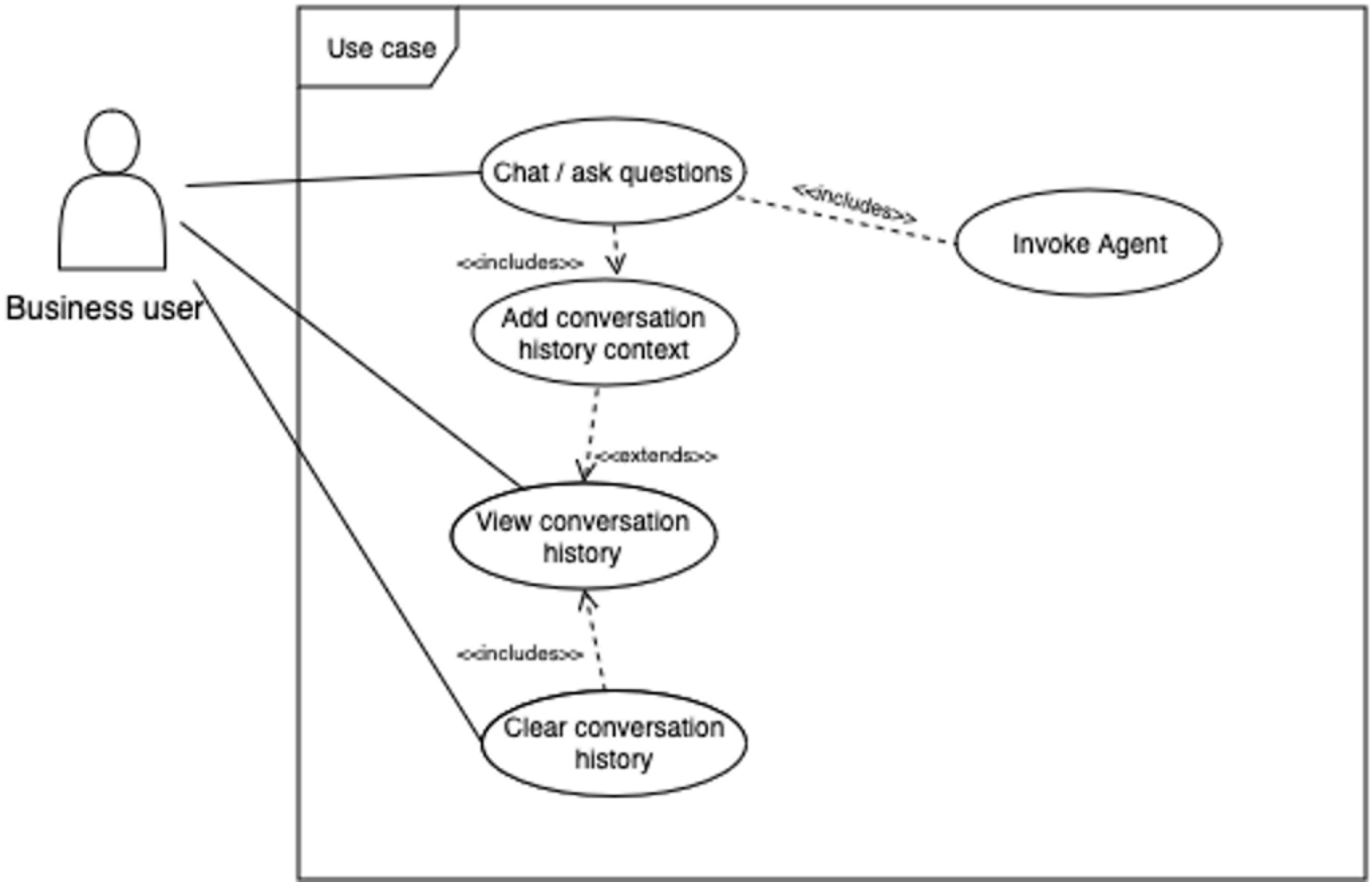
Représente le schéma de cas d'utilisation pour l'utilisateur professionnel du cas d'utilisation du texte



Avec le cas d'utilisation de Bedrock Agent, l'utilisateur professionnel peut accéder à une interface utilisateur pour discuter avec l'agent Amazon Bedrock configuré. L'utilisateur administrateur peut configurer ces spécifications dans les paramètres de déploiement. Dans le cas d'utilisation de Bedrock Agent, l'utilisateur professionnel a accès aux actions suivantes :

- Envoyer des messages via l'interface de chat
- Afficher l'historique des conversations
- Effacer l'historique des conversations

Représente le schéma de cas d'utilisation pour l'utilisateur professionnel du cas d'utilisation de Bedrock Agent



Générateur d'agents

L'Agent Builder fournit une plate-forme pour créer, déployer et gérer des agents d'IA prêts à être utilisés en production sur Amazon Bedrock. AgentCore Cette section décrit les composants techniques et les détails de mise en œuvre.

AgentCore intégration

Agent Builder utilise une approche de déploiement basée sur la configuration avec des images d'agent prédéfinies pour permettre des déploiements d'agents rapides, sécurisés et évolutifs.

Images d'agent prédéfinies

Les images des conteneurs d'agents sont créées par l'équipe GAAB pendant le CI/CD pipeline et publiées dans un référentiel ECR public. Chaque version d'image est liée à la version de la solution

GAAB (par exemple, v4.0.0 →:v4.0.0). gaab-strands-agent Les images sont basées sur le SDK Strands et incluent :

- Environnement d'exécution de l'agent
- Intégration du client MCP
- Capacités de gestion de mémoire
- OpenTelemetry instrumentation

Cache pull-through ECR

La solution utilise le cache ECR Pull-Through pour distribuer automatiquement les images des agents depuis le référentiel ECR public vers l'ECR privé du client. Ce service géré par AWS :

- Met en cache les images lors de la première extraction (délai de 2 à 5 minutes)
- Élimine la logique de copie d'images personnalisée
- Assure la disponibilité des images locales pour les déploiements ultérieurs
- Crée des règles de cache uniques par déploiement pour éviter les conflits

Stockage des configurations

Les configurations d'agent sont stockées dans DynamoDB à côté des configurations de cas d'utilisation existantes. Chaque configuration inclut :

- Modèle d'invite du système
- Fournisseur du modèle et identifiant du modèle
- Paramètres du modèle (température, max_tokens)
- Références et points de terminaison du serveur MCP
- Paramètres de mémoire (bascule de mémoire à long terme)
- Métadonnées de déploiement

Registre des versions d'images

Un tableau DynamoDB suit les versions d'image d'agent disponibles et leur URIs cache, permettant ainsi la gestion des versions et la rétrocompatibilité.

Configuration de l'agent

Invitations du système

Les instructions du système définissent le comportement, la personnalité et les capacités des agents. Les utilisateurs administrateurs peuvent :

- Modifiez le modèle par défaut via l'interface utilisateur d'Agent Builder
- Incluez des instructions pour l'utilisation de l'outil et le formatage des réponses
- Réinitialisez le modèle par défaut à tout moment

Sélection du modèle

Agent Builder prend en charge les modèles Amazon Bedrock dans la version 4.0.0 :

- Fournisseur de modèles : Amazon Bedrock (seule option dans la version 4.0.0)
- Sélection de modèles : Claude, Nova et autres modèles Bedrock
- Paramètres du modèle : température, max_tokens, top_p et paramètres spécifiques au modèle

Intégration au serveur MCP

Les serveurs Model Context Protocol permettent aux agents d'accéder aux outils et aux données de l'entreprise :

- Découverte du serveur via le point de terminaison de l'API GET/mcp
- Configuration dynamique sans modification du code
- Authentification et gestion des terminaux
- Exposition aux agents des capacités de l'outil

Streaming et traitement

Streaming en temps réel

Agent Builder utilise les événements envoyés par le serveur (SSE), du AgentCore bridge au WebSocket streaming des réponses en temps réel :

- La fonction Lambda établit une connexion SSE à Runtime AgentCore

- Les flux sont reliés à API Gateway WebSocket
- Permet de token-by-token fournir des réponses aux clients
- Maintient la connexion pour les demandes de longue durée

Contraintes de traitement

Le traitement de l'agent dans la version 4.0.0 est limité au délai d'exécution de Lambda :

- Temps de traitement maximal : 15 minutes
- Modèle de traitement synchrone
- Convient aux agents conversationnels et aux flux de travail modérés
- Support asynchrone étendu prévu pour la version 4.1 et les versions ultérieures

Gestion de mémoire

Mémoire à court terme

Activé par défaut pour tous les agents utilisant une option personnalisée MemoryHookProvider :

- Capture les événements de conversation via les gestionnaires de rappel Strands
- Organisé par ActorID et SessionId pour isoler le contexte
- Maintient le contexte des conversations au cours des sessions
- Intégration automatique à AgentCore la mémoire

Mémoire à long terme

Fonctionnalité optionnelle utilisant AgentCore Memory Tool de strands_tools :

- Basculement simple dans l'interface utilisateur d'Agent Builder
- Stratégie de mémoire sémantique avec paramètres par défaut
- Accès contrôlé par l'agent grâce à l'invocation naturelle des outils
- Stocke les informations extraites au fil des sessions
- Utilise ConversationID comme SessionId

Observabilité

OpenTelemetry Distribution AWS (ADOT)

Les agents sont automatiquement instrumentés lors de la construction du conteneur :

- Génération automatique de traces pour les opérations des agents
- Traçage distribué au-delà des limites des services
- Journalisation structurée avec corrélation IDs
- Intégration à CloudWatch Transaction Search

Flux d'authentification

Les utilisateurs s'authentifient via Amazon Cognito à l'aide de jetons JWT validés par des autorisateurs Lambda personnalisés qui récupèrent les politiques IAM de DynamoDB en fonction des groupes d'utilisateurs.

Générateur de flux de travail

Workflow Builder permet une orchestration multi-agents en créant un agent superviseur qui coordonne plusieurs agents Agent Builder à l'aide du modèle de délégation Agents as Tools.

Architecture du flux de travail

Composantes clés

- Agent superviseur : agent Entrypoint qui reçoit les demandes des utilisateurs et délègue les tâches à des agents spécialisés
- Agents spécialisés : cas d'utilisation d'Agent Builder enregistrés sous forme d'outils pour le superviseur
- Registre des agents : table DynamoDB stockant les configurations et les métadonnées des agents
- Couche d'orchestration : implémentation du modèle Agents as Tools dans le SDK Strands

Instanciation de l'agent

Création d'un agent local

Tous les agents spécialisés sont instanciés localement dans le même AgentCore environnement d'exécution :

1. Récupère les configurations des agents depuis DynamoDB
2. Crée des instances locales de chaque agent Agent Builder
3. Chaque agent gère ses propres connexions au serveur MCP
4. L'agent superviseur enregistre les agents spécialisés en tant qu'outils
5. Le SDK Strands gère la sélection et la délégation des agents

Planifiez votre déploiement

Cette section décrit les considérations relatives aux coûts, à la sécurité, à la région et aux quotas pour planifier votre déploiement.

> **Important**
>
> Cette solution utilise Amazon Bedrock comme principal service d'accès aux modèles générés par l'IA. Vous devez d'abord demander l'accès aux modèles avant qu'ils ne puissent être utilisés dans la solution. Pour plus de détails, reportez-vous à la section Accès aux modèles dans le guide de l'utilisateur d'Amazon Bedrock.

Régions AWS prises en charge

> **Important**
>
> Cette solution utilise en option les services Amazon Bedrock et Amazon Kendra, qui ne sont actuellement pas disponibles dans toutes les régions AWS. Vous devez lancer cette solution dans une région AWS où ces services sont disponibles. Pour connaître la disponibilité la plus récente des services AWS par région, consultez la liste des services régionaux AWS.

Generative AI Application Builder sur AWS est pris en charge dans les régions AWS suivantes :

Nom de la région	
USA Est (Ohio)	Canada (Centre)
USA Est (Virginie du Nord)	Europe (Francfort)
USA Ouest (Californie du Nord)	Europe (Irlande)
USA Ouest (Oregon)	Europe (Londres)
Asie-Pacifique (Mumbai)	Europe (Milan)
Asie-Pacifique (Séoul)	Europe (Paris)

Nom de la région	
Asie-Pacifique (Singapour)	Europe (Stockholm)
Asie-Pacifique (Sydney)	Middle East (Bahrain)
Asie-Pacifique (Tokyo)	Amérique du Sud (São Paulo)

> **Note**
>
> Si vous utilisez un modèle de base accessible en dehors d'AWS dans vos déploiements, vérifiez auprès du fournisseur du modèle dans quelles régions il est disponible. APIs S'ils ne APIs sont disponibles que dans certaines régions, vous risquez de rencontrer une instabilité sous la forme d'une latence élevée ou même de délais d'attente. Il est également important de consulter les équipes juridiques et de conformité de votre organisation pour évaluer les considérations liées au franchissement des frontières régionales par les données.

Cost

Avec cette solution AWS, vous ne payez que pour les ressources que vous utilisez, sans frais minimaux ni frais d'installation. Les utilisateurs paient pour le tableau de bord utilisé pour lancer les cas d'utilisation de l'IA générative et pour tous les cas d'utilisation déployés. Le coût des cas d'utilisation déployés dépend des configurations. Exemples de configurations :

1. Un tableau de bord de déploiement simple qui coûte environ 20 USD par mois.
2. Un exemple d'utilisation simple d'un chatbot prêt pour la production, déployé avec des paramètres par défaut et exécuté dans l'est des États-Unis (Virginie du Nord), alimenté par Amazon Bedrock sans accès aux documents, qui coûte également environ 200 dollars américains par mois.
3. Un système évolutif adapté à un cas d'utilisation d'Amazon VPC qui prend en charge 8 000 requêtes par jour sur des dizaines de milliers de documents, pour un coût d'environ 1 500 dollars américains par mois. Le coût du cas d'utilisation varie en fonction de la configuration, par exemple les cas d'utilisation de texte avec différents fournisseurs de modèles, avec ou sans activation de la génération augmentée de récupération (RAG), etc.

Description de la charge de travail	Coût estimé (USD/mois)
Exemple de coût pour le tableau de bord de déploiement	20 $/mois
Coûts d'échantillonnage pour une preuve de concept basée sur du texte (inclut le tableau de bord de déploiement et 1 cas d'utilisation du texte, environ 100 interacti ons par jour)	40 $/mois
Exemples de coûts pour un moteur de requêtes génératif basé sur l'IA hautement évolutif (Comprend un tableau de bord de déploiement, un cas d'utilisation textuel et un index Amazon Kendra pour RAG (jusqu'à 100 000 documents avec environ 8 000 requêtes par jour, avec le VPC activé)	1 500 $/mois
Exemples de coûts pour une preuve de concept basée sur un agent (Comprend un tableau de bord de déploiement, un cas d'utilisation d'un agent Bedrock avec les bases de connaissances Amazon Bedrock et Amazon Bedrock Guardrails activé, environ 100 interactions par jour)	840 $/mois
Exemples de coûts pour le serveur MCP (Comprend un tableau de bord de déploieme nt, un cas d'utilisation d'un serveur MCP avec méthode Gateway pour l'intégration Lambda, environ 100 appels d'outils par jour)	22 $/mois
Exemples de coûts pour Agent Builder	55 $/mois

Description de la charge de travail	Coût estimé (USD/mois)
(Comprend un tableau de bord de déploiement, un cas d'utilisation d'Agent Builder avec intégration MCP et activation de la mémoire à long terme, environ 100 interactions par jour)	
Exemples de coûts pour Workflow Builder (Comprend un tableau de bord de déploiement, 1 flux de travail avec 3 agents Agent Builder, environ 100 interactions par jour)	109 $/mois

> **Important**
>
> Ces exemples sont uniquement destinés à vous aider à estimer les coûts liés à vos charges de travail spécifiques. L'utilisation de différentes LLMs configurations ou de services AWS peut modifier vos coûts (par exemple, serverless/on-demand billing vs. provisioned/time - factured). Pour gérer les coûts, nous vous recommandons de créer un budget via AWS Cost Explorer. Les prix sont susceptibles d'être modifiés. Pour plus de détails, consultez la page Web de tarification de chaque service AWS utilisé dans cette solution.

Exemples de coûts liés à l'exécution du tableau de bord de déploiement

Le tableau suivant fournit la répartition des coûts pour un tableau de bord de déploiement avec des paramètres par défaut et 100 utilisateurs actifs dans la région de l'est des États-Unis (Virginie du Nord) pendant un mois, ce qui coûtera environ 20 dollars par mois.

Service AWS	Dimensions	Coût [USD]
API Gateway, DynamoDB, CloudFront Amazon S3, Lambda, magasin de paramètres Systems Manager	5 000 appels d'API REST de 512 Ko par mois sans activation de la mise en cache	1,97$

Service AWS	Dimensions	Coût [USD]
Amazon Cognito	100 utilisateurs actifs par mois avec des fonctionnalités de sécurité avancées activées et aucun utilisateur ne se connectant via la fédération SAML ou OIDC	5,55$
AWS WAF	10 000 requêtes Web réparties sur 1 ACL Web et 7 règles définies sans aucun groupe de règles	12,60$
Coût total du tableau de bord de déploiement		20,12$

Coûts d'échantillonnage pour une preuve de concept basée sur du texte

Un tableau de bord de déploiement peut comporter de nombreux cas d'utilisation déployés à un moment donné. Le tableau suivant montre la répartition des coûts d'un cas d'utilisation déployé sans RAG pour 1 utilisateur professionnel effectuant 100 requêtes par jour avec le LLM. Les requêtes sont envoyées sous forme de message texte sur le WebSocket et la réponse est retransmise sous forme de jetons, en supposant que le streaming est activé. En utilisant le modèle Amazon Bedrock Nova Pro, le coût d'exécution de ce cas d'utilisation est d'environ 20 $/mois.

Service AWS	Dimensions	Coût [USD]
API Gateway (WebSocket) CloudFront, Lambda, Amazon S3, AWS Systems Manager Parameter Store	100 interactions par chat par jour. Taille moyenne des messages 32 Ko par message et 5 minutes par connexion.	0,61$
CloudWatch	CloudWatch Journaux de 1,5 Go avec mode détaillé activé à des fins d'expérimentation	7,23$

Service AWS	Dimensions	Coût [USD]
Amazon DynamoDB	Tableau de l'historique des conversations, 1 Go de stockage Table de configuration LLM, 1 Go de stockage	3,05$
Sous-total des coûts des cas d'utilisation (non compris LLMs)		10,89$
Amazon Bedrock (Nova Pro)	Hypothèses pour 100 interactions par jour : * Coût mensuel pour 190 000 jetons d'entrée par jour = 0,152$ × 30$ * Coût mensuel pour 16 000 jetons de sortie par jour = 0,0512 × 30$	6,10$
Coût total de l'application avec Amazon Bedrock (Nova Pro)	10,89$ (coût du cas d'utilisation) + 6,10$ (coût Amazon Bedrock)	17,00$

Note

Les coûts des appels d'inférence effectués vers des services extérieurs au réseau AWS ne sont pas inclus dans ces estimations. Reportez-vous au guide de tarification de votre fournisseur de LLM si vous n'utilisez pas de fournisseur de modèles AWS.
Les guides de tarification des services AWS sont disponibles à l'adresse suivante : tarification d'Amazon Bedrock et tarification d'Amazon SageMaker AI.

Exemples de coûts pour un moteur de requêtes génératif basé sur l'IA hautement évolutif

Le tableau suivant fournit la ventilation des coûts d'un cas d'utilisation compatible RAG avec le modèle Nova Pro d'Amazon Bedrock comme LLM. Lorsqu'une base de connaissances Bedrock est ajoutée, ce cas d'utilisation coûte environ 1300 $/mois

Service AWS	Dimensions	Coût [USD]
API Gateway (WebSocket)	8000 interactions par chat par jour. Taille moyenne des messages 32 Ko par message et 5 minutes par connexion.	38,89$
CloudFront	240 000 demandes par mois avec 100 Go de données transférées vers Internet et 1 Go de données transférées vers l'origine	8,76$
Amazon Bedrock (Nova Pro)	Hypothèses : Jetons d'entrée = PromptTemplate (400) + contexte (400) + ChatHistory (1080) + requête Jetons d'entrée (20) = 1 900 Jetons de sortie = 160 (moyenne) Avec 8 000 transactions par jour, Coût quotidien des jetons d'entrée (1 900 x 8 000 = 15 200 000 jetons x 0,0008/1000 prix par jeton)	487,80$

Service AWS	Dimensions	Coût [USD]
	Coût quotidien des jetons de sortie (160 x 8 000 = 1 280 000 jetons x 0,0032/1000 prix par jeton) Coût mensuel ((12,16$ + 4,10$) x 30)	
CloudWatch	24 métriques utilisant 5 Go de données ingérées pour les journaux et 1 tableau de bord	9,72$
DynamoDB	Table DynamoDB pour suivre l'historique des conversations avec chaque enregistrement jusqu'à 1 Ko de données, 8 000 lectures et écritures par jour	11,70$
Lambda	Taille du conteneur : 128 Mo, 512 Mo éphémère stockage, 2 fonctions Lambda utilisées pour l'autorisation Taille du conteneur : 256 Mo, 512 Mo de stockage éphémère, 5 demandes par seconde avec un temps de calcul moyen de 20 secondes	20,89$
Coût total des cas d'utilisation		577,76 $/mois + coût de la base de connaissances (voir ci-dessous)

> **Note**
>
> Les coûts des appels d'API effectués vers des services extérieurs au réseau AWS ne sont pas inclus dans ces estimations. Consultez le guide tarifaire de votre fournisseur de LLM si vous n'utilisez pas Amazon Bedrock.

Coûts liés à l'ajout d'une base de connaissances

Les coûts de la base de connaissances varieront en fonction du type de base de connaissances utilisé et (dans le cas de Bedrock) du magasin de vecteurs sous-jacent utilisé par la base de connaissances. Le provisionnement et la gestion des bases de connaissances n'entrent pas dans le cadre de la solution.

Bases de connaissances Amazon Bedrock

La solution ne gère ni ne fournit de ressources liées aux bases de connaissances Amazon Bedrock. Amazon Bedrock ne facture aucun frais pour l'utilisation de la fonctionnalité de base de connaissances elle-même, mais l'utilisation du modèle d'intégration utilisé dans votre cas d'utilisation vous sera facturée pour chaque requête. En outre, le magasin vectoriel sous-jacent à votre base de connaissances (par exemple, un index dans Amazon OpenSearch Service ou une base de données dans Amazon Relational Database Service) aura un coût associé qui ne peut être ni fourni ni calculé ici.

Pour le scénario de moteur de requêtes d'IA génératif hautement évolutif ci-dessus, les coûts engagés par ce service pour appeler le modèle d'intégration Amazon Bedrock sont les suivants :

Service AWS	Dimensions	Coût [USD]
Amazon Bedrock (Amazon Titan Text Embeddings V2)	8 000 requêtes par jour avec 1 900 jetons d'entrée par requête = 15 200 000 jetons = 0,30 USD par jour. Coût quotidien x 30 jours = coût mensuel de 9,00 USD	9,00$

Service AWS	Dimensions	Coût [USD]
Exemple d'utilisation d'Amazon OpenSearch Service (sans serveur)	Configuration sans serveur de base avec 4 unités de OpenSearch calcul (OCU) (minimum facturable) = 23,04 USD par jour Coût quotidien x 30 jours = 691,20$ USD **Note** Cela fournit une estimation approximative, car certaines charges de travail nécessiteront davantage OCUs, tandis que les clients disposant de OpenSearch ressources provisionnées auront à supporter des coûts moindres dans ce domaine.	691,20$
Coût supplémentaire total		700,20$

Amazon Kendra

La solution peut vous fournir un index Kendra, ou vous pouvez apporter le vôtre. Le coût d'exécution d'une configuration adaptée au moteur de requêtes IA génératif hautement évolutif ci-dessus est le suivant :

Service AWS	Dimensions	Coût [USD]
Amazon Kendra	0 à 8 000 requêtes par jour et jusqu'à 100 000 documents avec Amazon Kendra Enterprise Edition avec 0 à 50 sources de données	1 008,00$

> **Note**
>
> Vous pouvez partager l'index Amazon Kendra entre différents cas d'utilisation, mais cela peut augmenter le nombre de requêtes par index. S'il ne s'agit pas de l'édition Amazon Kendra Enterprise, des frais supplémentaires s'appliqueront.

Coût supplémentaire lié à l'activation d'Amazon VPC pour un cas d'utilisation

Le tableau suivant fournit la répartition des coûts liés à l'activation d'Amazon VPC pour un cas d'utilisation déployé en deux. AZs

Service AWS	Dimensions	Coût [USD]
Passerelle Amazon NAT	Hypothèse : déploiement de 2 zones, avec une passerelle NAT dans chaque zone de disponibilité. 100 Go de données traitées via NAT Gateway 730 heures, 100 Go de données traitées par mois	74,70$
AWS PrivateLink (points de terminaison VPC)	Hypothèses : déploiement à 2 AZ, avec 1 sous-réseau privé dans chaque AZ et 1 point de terminaison VPC	97,84$

Service AWS	Dimensions	Coût [USD]
	avec 2 interfaces ENIs réseau élastiques (). 6 points de terminaison VPC, 2 par point de terminaison ENIs VPC, 730 heures avec 1 024 Go de données traitées en un mois	
IPv4 Adresse publique	Hypothèse : déploiement de 2 AZ, 1 sous-réseau public dans chaque AZ avec une passerelle NAT dans chaque sous-réseau public. Chaque passerelle NAT est configurée avec 1 public actif IPv4. 2 IPv4 adresses publiques actives x 730 heures par mois x 0,005$ de frais horaires = 7,3 USD	7,30$
Coût supplémentaire (pour Amazon VPC)		179,93$

Incidences financières liées à l'utilisation du débit provisionné

Les coûts de débit provisionnés varient en fonction du type de modèle que vous avez provisionné et de votre période d'engagement, ainsi que des unités de modèle sélectionnées pour la période d'engagement. L'utilisation du débit provisionné entraîne un coût supplémentaire.

Pour plus d'informations et pour up-to-date connaître les tarifs les plus élevés, vous pouvez consulter les tarifs de Bedrock.

Coût de l'utilisation de l'inférence entre régions

L'utilisation de l'inférence entre régions n'entraîne aucun coût supplémentaire pour le routage ou le transfert de données. Vous payez le même prix par jeton pour les modèles que dans votre région source ou principale.

Exemples de coûts pour une preuve de concept basée sur un agent

Lorsque vous utilisez Amazon Bedrock Agents, vous êtes facturé en fonction des composants composant l'agent, tels que le modèle de support et la base de connaissances (si RAG est activé), ainsi que des fonctionnalités supplémentaires que vous ajoutez. Le tableau suivant présente la répartition des coûts d'un cas d'utilisation de Bedrock Agent configuré avec un modèle Claude 3.5 Sonnet à la demande, les bases de connaissances Amazon Bedrock et Amazon Bedrock Guardrails.

Tout comme le coût d'ajout des bases de connaissances Amazon Bedrock, cette solution ne gère ni ne fournit les ressources liées aux agents Amazon Bedrock. La solution n'entraîne pas non plus de frais pour l'utilisation des bases de connaissances Amazon Bedrock, mais entraîne des frais pour :

- Utilisation du modèle d'intégration pour chaque requête qui lui est envoyée
- Le magasin vectoriel sous-jacent à votre base de connaissances (par exemple, un index dans Amazon OpenSearch Service ou une base de données dans Amazon RDS)

Le tableau suivant suppose 100 interactions par jour avec 1 900 jetons d'entrée et 160 jetons de sortie par requête.

> **Note**
>
> Pour cet exemple de cas d'utilisation de l'agent Bedrock, si un groupe d'action était configuré pour utiliser une API externe, ces coûts seraient supplémentaires. Ils n'entrent pas dans le cadre des calculs présentés dans ce tableau.

Service AWS	Dimensions	Coût [USD]
API Gateway (WebSocket) CloudFront, Lambda, Amazon	100 interactions par chat par jour, taille moyenne des messages 32 Ko par	0,61$

Service AWS	Dimensions	Coût [USD]
S3, magasin de paramètres Systems Manager	message, 5 minutes par connexion	
CloudWatch	CloudWatch Logs de 1,5 Go avec mode détaillé activé à des fins d'expérimentation	7,23$
DynamoDB	Table de configuration LLM pour une taille d'enregis trement de 1 Ko et un stockage de 1 Go	0,25$
Sous-total des coûts (non compris LLMs)		8,09$
Sonnet Anthropic Claude 3.5	* Coût quotidien pour 190 000 jetons d'entrée par jour (0,003 /1 000 jetons) = 0,57$ + Coût quotidien × 30 jours = 17,10 $* Coût quotidien pour 16 000 jetons de sortie par jour (0,015 $/1 000 jetons) = 0,24$ + Coût quotidien × 30 jours = 7,20$	24,30$
Amazon Bedrock (Amazon Titan Text Embeddings V2) pour les bases de connaissa nces Amazon Bedrock	Coût quotidien pour 190 000 jetons d'entrée par jour (0,00002/1 000 jetons) = 0,004 Coût quotidien × 30 jours = 0,12$	0,12$

Service AWS	Dimensions	Coût [USD]
Exemple d'utilisation d'Amazon OpenSearch Service (sans serveur)	Configuration sans serveur de base avec 4 unités de OpenSearch calcul (OCU) (minimum facturable) = 23,04$ par jour Coût quotidien × 30 jours = 691,20$	691,20$
Barrières de protections Amazon Bedrock	190 000 jetons équivalent à peu près 760 000 caractères (190 000 × 4) et 3 800 unités de texte (760 000 caractères pour 200) Envisagez un garde-corps configuré avec des filtres de contenu, un filtre d'informa tions personnelles (PII), un filtre d'informations sensibles (expression régulière) et des filtres de mots Coût quotidien du filtre de contenu (0,75/1 000 unités de texte) + coût du filtre PII (0,1/1 000 unités de texte) + filtre d'informations sensibles (regex) + filtres de mots = 2,85$ + 0,38$ + 0$ Coût mensuel = coût quotidien × 30 jours = 96,90$	96,90$

Service AWS	Dimensions	Coût [USD]
Coût total de candidature pour un agent soutenu par Anthropic Claude 3.5 Sonnet	8,09$ (coût du cas d'utilisation) + 812,52$ (autres configurations d'agent)	820,61$

> **Note**
>
> Reportez-vous au guide de tarification de votre fournisseur de LLM si vous n'utilisez pas de fournisseur de modèles AWS. Les guides de tarification des services AWS sont disponibles à l'adresse suivante : tarification d'Amazon Bedrock et tarification d'Amazon SageMaker AI.

Exemples de coûts pour le serveur MCP

Les cas d'utilisation des serveurs MCP permettent le déploiement et la gestion de serveurs Model Context Protocol sur Amazon AgentCore Bedrock. Le tableau suivant montre la répartition des coûts d'un cas d'utilisation d'un serveur MCP utilisant la méthode Gateway pour encapsuler les fonctions Lambda existantes.

La solution gère le déploiement et la configuration de la AgentCore passerelle. Vous êtes facturé pour :

- Coûts d'infrastructure (API Gateway, Lambda, DynamoDB, S3) CloudWatch
- AgentCore Consommation de la passerelle (par appel d'outil)
- Coûts d'exécution de la fonction Lambda (pour la méthode Gateway avec des cibles Lambda)
- Coûts d'API externes (pour la méthode Gateway avec des cibles d'API ou de serveur MCP, le cas échéant)

Élément	Calculs	Cost
Amazon API Gateway (API REST)	100 appels d'outils par jour × 30 jours = 3 000 demandes par mois	0,05 USD

Élément	Calculs	Cost
AWS Lambda (orchestration)	100 appels par jour × 30 jours × 1 seconde en moyenne × 512 Mo = 3 000 Go-secondes par mois	0,05 USD
Amazon DynamoDB	3 000 read/write demandes par mois + 1 Go de stockage	0,15$
Amazon CloudWatch	Surveillance et journalisation standard pour 3 000 invocations	1,00$
Amazon S3	Stockage de configuration et journaux (utilisation minimale)	0,25$
Passerelle Amazon Bedrock AgentCore	3 000 appels d'outils par mois	0,05 USD
Fonction Lambda cible	100 appels par jour × 30 jours × 0,5 seconde × 128 Mo = 1 500 Go de secondes par mois	0,25$
Coût mensuel total	1,75$ (infrastructure) + 0,05$ (passerelle) AgentCore	1,80$

Note

Les coûts varient en fonction de la méthode de déploiement (Gateway ou Runtime), des types de cibles et des modèles d'utilisation. Les déploiements de méthodes d' AgentCore exécution entraînent des frais d'exécution plutôt que des frais de passerelle. Les coûts d'API externes et les coûts d'hébergement de conteneurs personnalisés sont supplémentaires.

Exemples de coûts pour Agent Builder

Agent Builder vous permet de créer et de déployer des agents personnalisés sur Amazon Bedrock AgentCore. Le tableau suivant indique la répartition des coûts d'un cas d'utilisation d'Agent Builder configuré avec Claude 3.5 Sonnet, intégration du serveur MCP et activation de la mémoire à long terme.

La solution gère le déploiement et la configuration du AgentCore Runtime. Vous êtes facturé pour :

- Coûts d'infrastructure (API Gateway, Lambda, DynamoDB, S3) CloudWatch
- AgentCore Consommation d'exécution (heures de processeur et de mémoire basées sur le temps d'exécution réel de l'agent)
- Inférence du modèle de base (jetons d'entrée et de sortie)
- AgentCore Mémoire (événements à court terme et stockage/récupération à long terme)

Le tableau suivant suppose 100 interactions par jour avec 1 900 jetons d'entrée et 160 jetons de sortie par requête, avec un temps d'exécution moyen de l'agent de 5 secondes par interaction.

Service AWS	Dimensions	Coût [USD]
API Gateway (WebSocket) CloudFront, Lambda, Amazon S3, magasin de paramètres Systems Manager	100 interactions par chat par jour, taille moyenne des messages 32 Ko par message, 5 minutes par connexion	0,61$
CloudWatch	CloudWatch Logs de 1,5 Go avec mode détaillé activé à des fins d'expérimentation	7,23$
DynamoDB	Table de configuration LLM pour une taille d'enregis trement de 1 Ko et un stockage de 1 Go	0,25$
Sous-total des coûts d'infrast ructure		8,09$

Service AWS	Dimensions	Coût [USD]
Amazon Bedrock Runtime AgentCore	* Processeur : 1 vCPU × 5 secondes × 100 interactions = 125 vCPU- seconds/day = 0.140 vCPU-hours/day + Coût quotidien : 0,140 × 0,0895$ = 0,013$ + Coût mensuel : 0,013$ × 30 = 0,38$ * Mémoire : 512 Mo (0,5 Go) × 5 secondes × 100 interacti ons = 250 Go- seconds/day = 0.069 GB-hours/day + Coût quotidien : 0,069 × 0,00945$ = 0,0007$ + Coût mensuel : 0,0007$ × 30 = 0,02$	0,40$
Sonnet Anthropic Claude 3.5	* Coût quotidien pour 190 000 jetons d'entrée par jour (0,003 par 1 000 jetons) = 0,57$ + coût quotidien × 30 jours = 17,10$ * Coût quotidien pour 16 000 jetons de sortie par jour (0,015 $/1 000 jetons) = 0,24$ + coût quotidien × 30 jours = 7,20$	24,30$

Service AWS	Dimensions	Coût [USD]
Mémoire Amazon Bedrock AgentCore	* Mémoire à court terme : 100 nouveaux événements events/day × 0,25 $/1 000 = 0,025 $/jour + Coût mensuel : 0,025$ × 30 = 0,75$ * Stockage de mémoire à long terme (stratégie intégrée) : 100 enregistrements × 0,75 $/1 000 = 0,075 $/mois records/month * Récupération de la mémoire à long terme : 100 retrievals/day × 0,50 $/1 000 extractions = 0,05 $/jour + Coût mensuel : 0,05 × 30 = 1,50$	2,33$
Coût total de l'application pour Agent Builder avec Claude 3.5 Sonnet	8,09$ (infrastructure) + 0,40$ (temps AgentCore d'exécution) + 24,30$ (modèle) + 2,33$ (mémoire)	35,12$

Note

AgentCore La tarification du temps d'exécution est basée sur la consommation. Les coûts réels dépendent de :

- Durée d'exécution de l'agent (utilisation du processeur et de la mémoire pendant le traitement actif)
- Nombre d'interactions et leur complexité
- Utilisation de l'outil MCP (supplémentaire CPU/memory pour l'exécution de l'outil)
- Configuration de la mémoire (mémoire à court terme ou mémoire à long terme activée)

Pour connaître les AgentCore tarifs détaillés, consultez les tarifs d'Amazon Bedrock.

> **Note**
>
> Si vous utilisez des serveurs MCP qui font appel à des services APIs ou à des services externes, ces coûts sont supplémentaires et ne sont pas pris en compte dans ce calcul. De même, si vous utilisez des outils de AgentCore navigateur ou d'interprétation de code, des frais basés sur la consommation s'appliquent à 0,0895$ par heure de vCPU et à 0,00945$ par Go d'heure.

Exemples de coûts pour Workflow Builder

Workflow Builder crée un agent superviseur qui orchestre plusieurs agents Agent Builder. Le tableau suivant indique la répartition des coûts pour un flux de travail comprenant 1 agent superviseur et 3 agents Agent Builder spécialisés, tous configurés avec Claude 3.5 Sonnet et avec une mémoire à long terme activée.

Hypothèses : 100 interactions par jour, moyenne de 2 délégations d'agent par interaction, 5 secondes de temps d'exécution par agent.

Service AWS	Dimensions	Coût [USD]
API Gateway (WebSocket) CloudFront, Lambda, Amazon S3, magasin de paramètres Systems Manager	100 interactions par chat par jour, taille moyenne des messages 32 Ko par message, 5 minutes par connexion	0,61$
CloudWatch	CloudWatch Logs de 1,5 Go avec mode détaillé activé à des fins d'expérimentation	7,23$
DynamoDB	Table de configuration LLM pour une taille d'enregis	0,25$

Service AWS	Dimensions	Coût [USD]
	trement de 1 Ko et un stockage de 1 Go	
Sous-total des coûts d'infrast ructure		8,09$
Amazon Bedrock AgentCore Runtime (agent superviseur)	* Processeur : 1 vCPU × 5 secondes × 100 interactions = 0,140 vCPU hours/day × 30 = $0.38 * Memory: 0.5 GB × 5 seconds × 100 interactions = 0.069 GB-hours/day - × 30 = 0,02$	0,40$
Amazon Bedrock AgentCore Runtime (3 agents spécialisés)	* En moyenne, 2 délégations par interaction = 200 agents executions/day * CPU: 1 vCPU × 5 seconds × 200 = 0.278 vCPU-hours/day × 30 = $0.75 * Memory: 0.5 GB × 5 seconds × 200 = 0.139 GB-hours/day × 30 = 0,04$	0,79$
Anthropic Claude 3.5 Sonnet (agent superviseur)	* Entrée : 190 000$ tokens/da y × 0,003 $/1 000$ = 0,57 $/ jour × 30 = 17,10$ * Sortie : 16 000 × 0,015 $/1 000$ = 0,24$ par jour × 30 = 7,20$ tokens/ day	24,30$

Service AWS	Dimensions	Coût [USD]
Anthropic Claude 3.5 Sonnet (Agents spécialisés)	* En moyenne, 2 délégations par interaction * Entrée : 380 000 dollars × 0,003 $/1 000 dollars = 1,14 $/jour tokens/day × 30 = 34,20 dollars * Sortie : 32 000 × 0,015 $/1 000 dollars = 0,48 $/jour × 30 = 14,40 dollars tokens/day	48,60$
Amazon Bedrock AgentCore Memory (agent superviseur)	* Court terme : 100 events/day × 0,25 $/1 000 × 30 = 0,75$ * Stockage à long terme : 100 enregistrements × 0,75 $/1 000$ = 0,08$ * Récupération à long terme : 100 × 0,50 $/1 000 × 30 = 1,50$ retrievals/day	2,33$
Amazon Bedrock AgentCore Memory (agents spécialisés)	* Court terme : 200 events/day × 0,25 $/1 000 × 30 = 1,50$ * Stockage à long terme : 200 enregistrements × 0,75 $/1 000$ = 0,15$ * Récupération à long terme : 200 × 0,50 $/1 000 × 30 = 3$ retrievals/day	4,65$
Coût total de l'application pour Workflow Builder avec 3 agents	8,09$ (infrastructure) + 1,19$ (temps AgentCore d'exécution) + 72,90$ (modèles) + 6,98$ (mémoire)	89,16$

> **Note**
>
> - Des taux de délégation plus élevés augmentent proportionnellement la consommation de jetons
>
> Pour connaître les AgentCore tarifs détaillés, consultez les tarifs d'Amazon Bedrock.

Sécurité

Lorsque vous créez des systèmes sur l'infrastructure AWS, les responsabilités en matière de sécurité sont partagées entre vous et AWS. Ce modèle de responsabilité partagée réduit votre charge opérationnelle car AWS exploite, gère et contrôle les composants, notamment le système d'exploitation hôte, la couche de virtualisation et la sécurité physique des installations dans lesquelles les services fonctionnent. Pour plus d'informations sur la sécurité AWS, rendez-vous sur AWS Cloud Security.

Utilisation de modèles de fondation sur Amazon Bedrock

Amazon Bedrock héberge une collection de modèles allant des modèles Amazon Nova à d'autres modèles de base de premier plan (FMs). Lorsque vous utilisez Amazon Bedrock, tous les modèles sont hébergés au sein de l'infrastructure AWS. Cela signifie que lorsque vous utilisez Amazon Bedrock comme fournisseur de LLM, toutes vos demandes d'inférence resteront dans le réseau AWS et le trafic réseau ne quittera pas votre région.

> **Note**
>
> Tous les modèles de base (FMs) disponibles via Amazon Bedrock sont hébergés directement sur l'infrastructure AWS gérée et détenue par AWS. Les fournisseurs de modèles n'ont pas accès aux données des clients telles que les instructions et les continuations, ni aux journaux de service Amazon Bedrock. Pour plus d'informations sur le niveau de sécurité d'Amazon Bedrock, reportez-vous à la section Protection des données dans Amazon Bedrock dans le guide de l'utilisateur d'Amazon Bedrock.

Rôles IAM

Les rôles IAM permettent aux clients d'attribuer des politiques d'accès et des autorisations détaillées aux services et aux utilisateurs sur le cloud AWS. Cette solution crée des rôles IAM qui accordent aux fonctions Lambda de la solution l'accès pour créer des ressources régionales.

CloudWatch Journaux

Vous pouvez activer le mode détaillé lors du déploiement d'un cas d'utilisation à l'aide de la page de sélection du modèle du tableau de bord de déploiement, sous Paramètres supplémentaires. Le mode détaillé permet d'obtenir des CloudWatch journaux détaillés qui peuvent être utiles pour le débogage et une expérimentation rapide.

> **Note**
>
> Lorsque le mode détaillé est activé, les documents extraits de la base de connaissances (si RAG est activé) et les invites sont également enregistrés, ce qui peut contenir des informations sensibles.

VPC

La solution propose deux options pour la configuration d'Amazon VPC :

1. Laissez la solution créer un Amazon VPC pour vous.
2. Gérer et intégrer votre propre Amazon VPC à utiliser dans le cadre de la solution.

Laissez la solution créer un Amazon VPC pour vous

Si vous sélectionnez l'option permettant à la solution de créer un Amazon VPC, celle-ci sera déployée en tant qu'architecture 2-AZ par défaut avec une plage d'adresses CIDR 10.10.0.0/20. Vous avez la possibilité d'utiliser Amazon VPC IP Address Manager (IPAM), avec 1 sous-réseau public et 1 sous-réseau privé dans chaque zone de disponibilité. La solution crée des passerelles NAT dans chacun des sous-réseaux publics et configure les fonctions Lambda pour les créer dans ENIsles sous-réseaux privés. En outre, cette configuration crée des tables de routage et leurs entrées, des groupes de sécurité et leurs règles, un réseau ACLs, des points de terminaison VPC (points de terminaison de passerelle et d'interface).

Gérer votre propre Amazon VPC

Lorsque vous déployez la solution avec un Amazon VPC, vous avez la possibilité d'utiliser un Amazon VPC existant dans votre compte et votre région AWS. Nous vous recommandons de rendre votre VPC disponible dans au moins deux zones de disponibilité pour garantir une haute disponibilité. Votre VPC doit également disposer des points de terminaison VPC suivants et de leurs politiques IAM associées pour les configurations de votre VPC et de table de routage.

Pour un tableau de bord de déploiement : Amazon VPC

1. Point de terminaison de passerelle pour DynamoDB.
2. Point de terminaison Gateway pour S3.
3. Point de terminaison de l'interface pour CloudWatch.
4. Point de terminaison de l'interface pour AWS CloudFormation.

Pour un cas d'utilisation : Amazon VPC

1. Point de terminaison de passerelle pour DynamoDB.
2. Point de terminaison Gateway pour S3.
3. Point de terminaison de l'interface pour CloudWatch.
4. Point de terminaison de l'interface pour le magasin de paramètres Systems Manager.

> **Note**
>
> La solution ne nécessite que`com.amazonaws.region.ssm`.

5. Point de terminaison de l'interface pour Amazon Bedrock (bedrock-runtime, agent-runtime,). bedrock-agent-runtime
6. Facultatif : si le déploiement utilise Amazon Kendra comme base de connaissances, un point de terminaison d'interface pour Amazon Kendra est nécessaire.
7. Facultatif : si le déploiement utilise un LLM sous Amazon Bedrock, un point de terminaison d'interface pour Amazon Bedrock est nécessaire.

> **Note**
>
> La solution ne nécessite que`com.amazonaws.region.bedrock-runtime`.

8. Facultatif : si le déploiement utilise Amazon SageMaker AI pour le LLM, un point de terminaison d'interface pour Amazon SageMaker AI est nécessaire.

> **Note**
>
> La solution ne supprimera ni ne modifiera la configuration du VPC lors de l'utilisation de l'option de déploiement Bring your own VPC. Cependant, il supprimera tous VPCs ceux créés par la solution dans l'option Créer un VPC pour moi. Pour cette raison, vous devez être prudent lorsque vous partagez un VPC géré par une solution entre plusieurs stacks/déploiements.
> Par exemple, le déploiement A utilise l'option Create a VPC for me. Le déploiement B utilise Bring my own VPC en utilisant le VPC créé par le déploiement A. Si le déploiement A est supprimé avant le déploiement B, le déploiement B ne fonctionnera plus car le VPC a été supprimé. De plus, étant donné que le déploiement B utilise les fonctions ENIs créées par les fonctions Lambda, la suppression du déploiement A peut entraîner des erreurs et la rétention de ressources résiduelles.

Amazon CloudFront

Cette solution déploie une console Web hébergée dans un compartiment Amazon S3. Pour réduire le temps de latence et améliorer la sécurité, cette solution inclut une CloudFront distribution dotée d'une identité d'accès d'origine, c'est-à-dire un CloudFront utilisateur fournissant un accès public au contenu du bucket du site Web de la solution. Pour plus d'informations, consultez Restreindre l'accès au contenu Amazon S3 à l'aide d'une identité d'accès d'origine dans le manuel Amazon CloudFront Developer Guide.

> **Note**
>
> CloudFront dispose d'une limite de quota souple au niveau du compte de 20 politiques d'en-tête de réponse. Cette solution crée des politiques d'en-tête de réponse personnalisées pour des raisons de sécurité. Si vous avez déployé plus de 20 applications Generative AI Application Builder sur AWS ou dans ses cas d'utilisation, les nouveaux déploiements risquent d'échouer en raison de l'atteinte de la limite de quota.

Pour résoudre ce problème, vous pouvez demander une augmentation de quota pour le quota Response Header Policies dans la console AWS Service Quotas en suivant ces étapes :

1. Ouvrez la console AWS Service Quotas.
2. Dans le volet de navigation, sélectionnez Services AWS.
3. Recherchez et sélectionnez Amazon CloudFront.
4. Accédez au quota des politiques d'en-tête de réponse et choisissez Demander une augmentation du quota.
5. Suivez les instructions pour demander une augmentation de la limite de quota pour votre compte AWS.

En augmentant le quota des politiques d'en-tête de réponse, vous pouvez vous assurer que les nouveaux déploiements du générateur d'applications d'IA générative sur AWS ou ses cas d'utilisation n'échoueront pas en raison de la limite de quota.

Quotas

Les quotas de service, également appelés limites, représentent le nombre maximal de ressources ou d'opérations de service pour votre compte AWS.

Quotas pour les services AWS dans cette solution

Assurez-vous de disposer d'un quota suffisant pour chacun des services mis en œuvre dans cette solution. Pour plus d'informations, consultez la section Quotas de service AWS.

Utilisez les liens suivants pour accéder à la page de ce service. Pour consulter les quotas de service pour tous les services AWS dans la documentation sans changer de page, consultez plutôt les informations de la page Points de terminaison et quotas du service dans le PDF.

Quotas Amazon Bedrock AgentCore

Pour les déploiements d'Agent Builder, tenez compte des quotas de AgentCore service Amazon Bedrock suivants :

Quota	USA Est (Virginie du Nord)	Autres régions
Charges de travail Active Session par compte	1 000	500
Nombre total d'agents par compte	1 000	1 000
Versions par compte	1 000	1 000

Déploiement de la solution

Cette solution utilise des CloudFormation modèles et des piles AWS pour automatiser son déploiement. Le CloudFormation modèle indique les ressources AWS incluses dans cette solution et leurs propriétés. La CloudFormation pile fournit les ressources décrites dans le modèle.

Vue d'ensemble du processus de déploiement

Avant de lancer la solution, examinez le coût, l'architecture, la sécurité et les autres considérations abordées dans ce guide.

> **Important**
>
> Si vous prévoyez d'utiliser Amazon Bedrock, vous devez demander l'accès aux modèles avant qu'ils ne soient disponibles. Reportez-vous à la section Accès aux modèles dans le guide de l'utilisateur d'Amazon Bedrock pour plus de détails.

Temps de déploiement : environ 10 minutes

Étape 1 : Lancer la pile de tableaux de bord de déploiement

Étape 2 : Déployer un cas d'utilisation

Étape 3 : Déployer un cas d'utilisation à l'aide de l'assistant du tableau de bord de déploiement

Étape 4 : Configuration après le déploiement

Vous pouvez éventuellement déployer les cas d'utilisation séparément de la solution, si vous préférez ne pas avoir l'interface utilisateur du tableau de bord de déploiement ou APIs.

- Déploiement d'un cas d'utilisation de texte autonome
- Déploiement d'un cas d'utilisation autonome de l'agent Bedrock

Vous pouvez également fournir une configuration de chat DynamoDB.

> **Important**
>
> Cette solution envoie des métriques opérationnelles à AWS (les « données ») concernant l'utilisation de cette solution. Nous utilisons ces données pour mieux comprendre comment les clients utilisent cette solution et les services et produits associés. La collecte de ces données par AWS est soumise à la politique de confidentialité d'AWS.

CloudFormation Modèle AWS

Vous pouvez télécharger le CloudFormation modèle de cette solution avant de la déployer.

View template

generative-ai-application-builder-on-aws.template - Utilisez ce modèle pour lancer la solution et tous les composants associés. La configuration par défaut déploie les solutions de base et de support figurant dans les services AWS de cette section de solutions, mais vous pouvez personnaliser le modèle en fonction de vos besoins spécifiques.

> **Note**
>
> Les CloudFormation ressources AWS sont créées à partir des constructions du kit AWS Cloud Development Kit (AWS CDK).

Ce CloudFormation modèle AWS déploie Generative AI Application Builder sur AWS dans le cloud AWS.

Étape 1 : Lancer la pile de tableaux de bord de déploiement

Suivez les step-by-step instructions de cette section pour configurer et déployer la solution dans votre compte.

Temps de déploiement : environ 10 minutes

1. Connectez-vous à l'AWS Management Console et sélectionnez le bouton pour lancer le `generative-ai-application-builder-on-aws.template` CloudFormation modèle.

   Launch solution

2. Le modèle est lancé par défaut dans la région USA Est (Virginie du Nord). Pour lancer la solution dans une autre région AWS, utilisez le sélecteur de région dans la barre de navigation de la console.

   > **Note**
   >
   > Cette solution utilise Amazon Kendra et Amazon Bedrock, qui ne sont actuellement pas disponibles dans toutes les régions AWS. Si vous utilisez ces fonctionnalités, vous devez lancer cette solution dans une région AWS où ces services sont disponibles. Pour connaître la disponibilité la plus récente par région, consultez la liste des services régionaux AWS.

3. Sur la page Create stack, vérifiez que l'URL du modèle est correcte dans la zone de texte URL Amazon S3 et choisissez Next.
4. Sur la page Spécifier les détails de la pile, attribuez un nom à votre pile de solutions. Pour plus d'informations sur les limites relatives aux caractères de dénomination, consultez les limites IAM et STS dans le guide de l'utilisateur d'AWS Identity and Access Management.
5. Sous Paramètres, passez en revue les paramètres de ce modèle de solution et modifiez-les si nécessaire. Cette solution utilise les valeurs par défaut suivantes.

Paramètre	Par défaut	Description
Adresse e-mail de l'utilisateur administrateur	No	Adresse e-mail de l'utilisateur administrateur qui aura accès au tableau de bord de déploiement. S'ils sont fournis, un groupe et un utilisateur Amazon Cognito seront créés avec les autorisations nécessaires pour déployer et gérer les cas d'utilisation. Vous

Paramètre	Par défaut	Description
		pouvez également l'utiliser `placeholder@example.com` pour créer le groupe, mais pas l'utilisateur. Reportez-vous à la section Configuration manuelle du groupe d'utilisateurs pour plus d'informations sur la configuration de votre groupe d'utilisateurs.
VpcEnabled	No	Le tableau de bord de déploiement doit-il être déployé au sein d'un VPC
CreateNewVpc	No	Disponible uniquement, si VpcEnabledc'est le cas`Yes`. Si la valeur est`Yes`, la pile créera le VPC et déploiera la solution au sein du VPC créé. Si VpcEnabledc'est le cas `Yes` et si CreateNew Vpcc'est le cas`No`, vous devez fournir une configuration VPC existante (ExistingVpcId,, ExistingPrivateSubnetIdsExistingSecurityGroupIds, VpcAzs).

Paramètre	Par défaut	Description
IPAMPoolId	(Entrée facultative)	Vous pouvez configurer IPAM et fournir l'identifiant créé en entrée pour attribuer la plage d'adresses IP que le déploiement de cette pile doit utiliser. Pour en savoir plus sur l'IPAM, voir Fonctionnement de l'IPAM.
Interface utilisateur de déploiement	`Yes`	Vous avez la possibilité de déployer le tableau de bord de déploiement sans l'interface utilisateur Web (et les ressources AWS requises pour le déploiement Web). Dans ce cas, la solution déploiera toute l'infrastructure, y compris les points de terminaison de l'API REST. Cette option est utile pour intégrer votre propre interface Web au tableau de bord de déploiement APIs.
ExistingVpcId	(Entrée facultative)	Obligatoire uniquement si vous souhaitez déployer la solution dans un VPC existant que vous avez créé.

Paramètre	Par défaut	Description
ExistingPrivateSubnetIds	(Entrée facultative)	Obligatoire uniquement si vous souhaitez déployer la solution dans un VPC existant que vous avez créé. Les fonctions Lambda seront déployées dans ce sous-réseau.
ExistingSecurityGroupIds	(Entrée facultative)	Obligatoire uniquement si vous souhaitez déployer la solution dans un VPC existant que vous avez créé. Assurez-vous que les groupes de sécurité disposent des autorisations nécessaires pour une connexion TCP sortante.
VpcAzs	(Entrée facultative)	Obligatoire uniquement si vous souhaitez déployer la solution dans un VPC existant que vous avez créé.
CognitoDomainPrefix	(Entrée facultative)	Obligatoire uniquement si vous souhaitez déployer la solution dans un groupe d'utilisateurs Amazon Cognito existant que vous avez créé. Si vous ne fournissez aucune valeur, la solution la génère.

Paramètre	Par défaut	Description
ExistingCognitoUserPoolId	(Entrée facultative)	Obligatoire uniquement si vous souhaitez déployer la solution dans un groupe d'utilisateurs Amazon Cognito existant que vous avez créé.
ExistingCognitoUserPoolClient	(Entrée facultative)	Obligatoire uniquement si vous souhaitez déployer la solution dans un groupe d'utilisateurs Amazon Cognito existant que vous avez créé. Si vous ne fournissez aucune valeur, la solution crée un client de groupe d'utilisateurs. Ce paramètre ne peut être fourni que si vous fournissez une ExistingCognitoUserPoolIdvaleur.

6. Choisissez Next (Suivant).
7. Sur la page Configurer les options de pile, choisissez Suivant.
8. Sur la page Réviser et créer, vérifiez et confirmez les paramètres. Cochez la case indiquant que le modèle créera des ressources AWS Identity and Access Management (IAM).
9. Choisissez Submit pour déployer la pile.

Vous pouvez consulter l'état de la pile dans la CloudFormation console AWS dans la colonne Status. Vous devriez recevoir le statut CREATE_COMPLETE dans environ 10 minutes.

Étape 2 : Déployer un cas d'utilisation

> **Important**
>
> Une fois que la pile a été déployée avec succès, un e-mail d'inscription est envoyé à l'adresse e-mail de l'utilisateur administrateur configuré. À l'aide de ces informations d'identification,

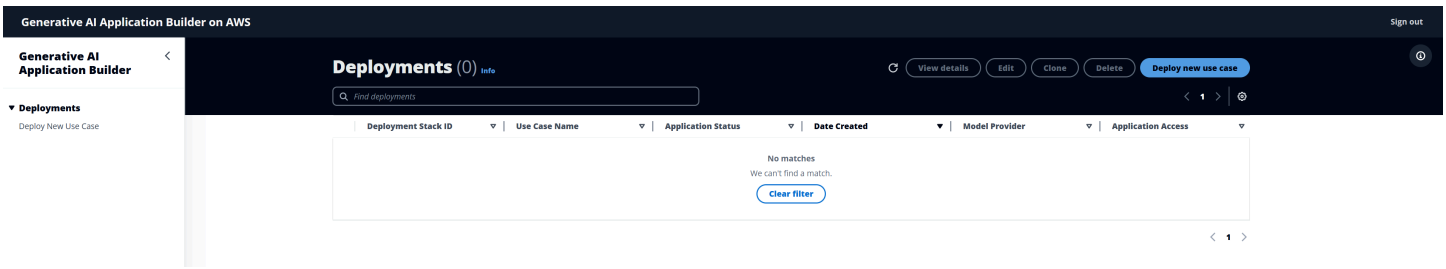
l'utilisateur administrateur peut se connecter au tableau de bord de déploiement pour utiliser l'application Web.

> **Note**
>
> L' DevOps utilisateur ayant accès à l'AWS Management Console doit fournir à l'administrateur l' CloudFront URL de l'interface utilisateur du tableau de bord de déploiement lorsque la pile est terminée. L'URL se trouve dans l'onglet Sorties de la CloudFormation pile.

1. Connectez-vous au tableau de bord de déploiement en tant qu'utilisateur administrateur.
2. Sur la page d'accueil de l'application, choisissez Déployer un nouveau cas d'utilisation.

   Cela lance l'assistant de déploiement, qui vous guide tout au long de la création du cas d'utilisation.

Représente la page d'accueil du tableau de bord de déploiement : nouveau déploiement



> **Note**
>
> Si vous devez ajouter des utilisateurs supplémentaires à votre déploiement, reportez-vous à la section Gestion du groupe d'utilisateurs de Cognito pour plus de détails.

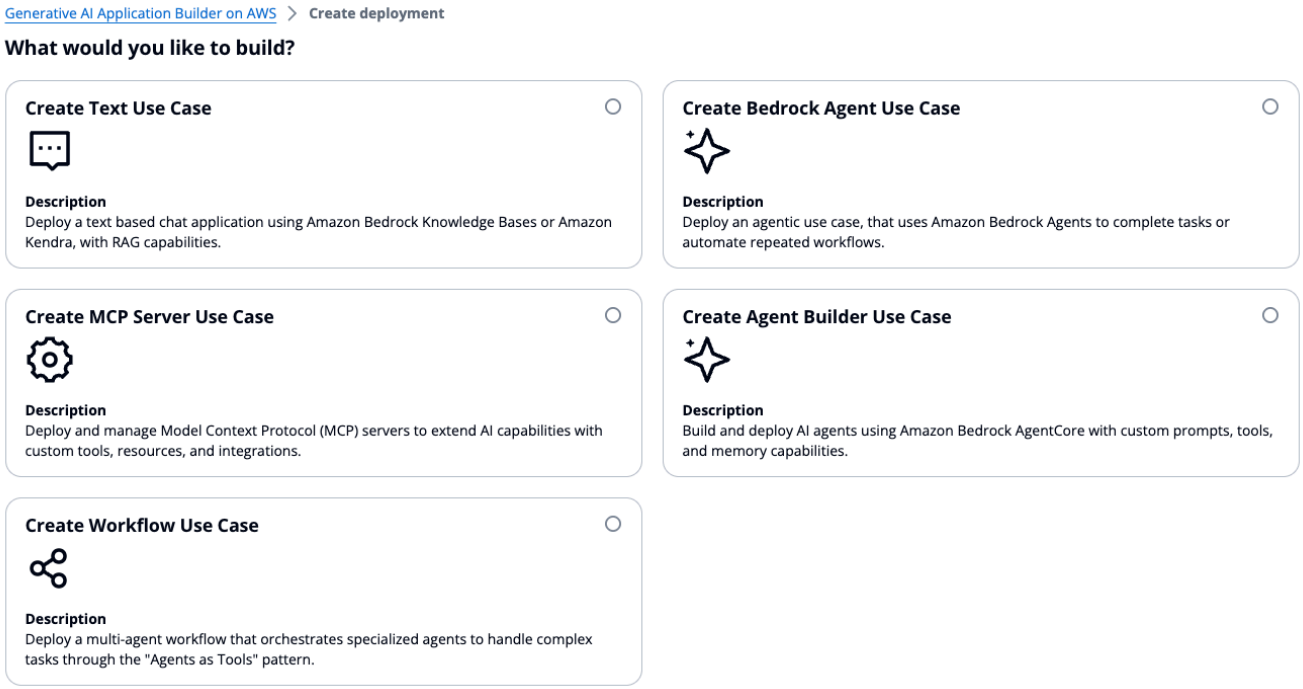
Étape 3 : Déployer un cas d'utilisation à l'aide de l'assistant du tableau de bord de déploiement

Dans l'assistant du tableau de bord de déploiement, vous devez choisir entre les options suivantes :

- Cas d'utilisation du texte - Déploie une application de chat, avec des fonctionnalités RAG en option

- Cas d'utilisation de Bedrock Agent - Utilise les agents Amazon Bedrock pour effectuer des tâches ou automatiser des flux de travail répétés
- Serveur MCP - Déployez et gérez des serveurs MCP à l'aide de passerelles ou de méthodes d'exécution
- Agent Builder : créez et déployez des agents personnalisés AgentCore grâce à l'intégration de MCP et à la gestion de la mémoire
- Générateur de flux de travail : orchestrez plusieurs agents Agent Builder à l'aide de la délégation hiérarchique

Affiche cinq options : créer un cas d'utilisation du texte, créer un cas d'utilisation de l'agent Bedrock, créer un cas d'utilisation du serveur MCP, créer un cas d'utilisation du générateur d'agents ou créer un cas d'utilisation du flux de travail.



Étape 3a : Déployer un cas d'utilisation de texte

Cette section fournit des instructions pour déployer un cas d'utilisation de type Text.

Sélectionnez un cas d'utilisation

Lorsque vous choisissez Créer un cas d'utilisation du texte, l'interface utilisateur ouvre l'écran Sélectionner un cas d'utilisation. Saisissez les informations suivantes :

- Nom du cas d'utilisation.
- Adresse e-mail facultative pour que l'utilisateur par défaut du cas d'utilisation soit ajouté au groupe d'utilisateurs Amazon Cognito pour le cas d'utilisation et soit autorisé à interagir avec celui-ci.
- Si vous souhaitez déployer une interface utilisateur avec ce cas d'utilisation. Si vous ne souhaitez pas déployer d'interface utilisateur avec le cas d'utilisation, vous pouvez utiliser les points de terminaison d'API déployés pour les utiliser avec votre application.

Détails du cas d’utilisation

L'étape des détails du cas d'utilisation vous permet de configurer des paramètres supplémentaires pour votre déploiement.

Par défaut, le cas d'utilisation du texte crée et configure un groupe d'utilisateurs Amazon Cognito pour vous lorsque la solution déploie le tableau de bord de déploiement. La solution authentifie les nouveaux cas d'utilisation auprès d'un client nouvellement créé dans le même groupe d'utilisateurs. Toutefois, vous pouvez fournir un identifiant de groupe d'utilisateurs et un identifiant client existants au cours de cette étape si vous souhaitez utiliser votre propre groupe d'utilisateurs et votre propre client Amazon Cognito dans le cadre de ce cas d'utilisation.

Important

Les utilisateurs administrateurs ont accès à tous les cas d'utilisation déployés lorsque le groupe d'utilisateurs Amazon Cognito est créé via l'assistant de déploiement. Si vous fournissez votre propre groupe d'utilisateurs lors du déploiement, vous devez vous assurer que l'administrateur dispose des autorisations nécessaires pour accéder aux cas d'utilisation déployés.
Vous devrez également mettre à jour le rappel autorisé URLs et la déconnexion autorisée dans les clients de votre application URLs dans Cognito. Pour cela :

1. Accédez à la console Cognito
2. Choisissez Groupes d’utilisateurs.
3. Choisissez votre groupe d'utilisateurs.
4. Choisissez App Clients dans le menu de gauche.
5. Choisissez le client d'application que vous souhaitez modifier.
6. Choisissez l'onglet Pages de connexion.
7. Choisissez Modifier et ajoutez votre URLs.

8. Sélectionnez Enregistrer les modifications.

En outre, si vous devez ajouter d'autres utilisateurs à un cas d'utilisation, consultez la section Gestion du groupe d'utilisateurs de Cognito.

Sélectionnez la configuration réseau

Cette étape de l'assistant vous permet de déployer le cas d'utilisation avec un Amazon Virtual Private Cloud (Amazon VPC) existant ou nouveau. Si vous sélectionnez un VPC préexistant, vous devez fournir un ID de VPC, jusqu'à 16 identifiants de sous-réseau et jusqu'à 5 groupes de sécurité IDs à utiliser avec ce VPC. Si vous n'utilisez pas de VPC préexistant, ces paramètres seront configurés pour vous.

Sélection d’un modèle

À l'étape Sélectionner un modèle, vous pouvez choisir votre fournisseur de modèles dans le menu déroulant. Il existe deux options : Bedrock et. SageMaker

Si vous le sélectionnez SageMaker, vous pouvez créer un point de terminaison du modèle SageMaker SageMaker AI dans la console AI et fournir le schéma d'entrée que le modèle attend et produit JSONPath pour la réponse LLM. Vous pouvez vous référer à la section Utiliser Amazon SageMaker AI en tant que fournisseur de LLM et aux exemples de charge utile d'SageMaker IA fournis dans le référentiel de GitHub la solution.

Si vous sélectionnez Amazon Bedrock, quatre options vous seront proposées :

- Modèles à démarrage rapide - Démarrez rapidement avec une collection de modèles aux price/performance caractéristiques différentes. Recommandé pour créer vos premières applications. Cette option vous permet de sélectionner un nom de modèle dans la liste fournie.
- Autres modèles de fondation - Accédez à la gamme complète de modèles de base avec différentes capacités et spécialisations. Cette option vous permet de saisir l'identifiant du modèle de fondation Bedrock à la demande que vous souhaitez.
- Profils d'inférence : les profils d'inférence tirent parti de l'inférence interrégionale de Bedrock pour augmenter le débit et améliorer la résilience en acheminant vos demandes entre plusieurs régions AWS pendant les pics d'utilisation. Cette option vous permet de saisir l'ID du profil d'inférence que vous souhaitez utiliser.

- Modèles provisionnés : capacité de débit dédiée aux charges de travail de production nécessitant des performances constantes. Cette option vous permet de saisir l'ARN du provisioned/custom modèle à utiliser depuis Amazon Bedrock.

L'étape de sélection du modèle vous permet également de choisir les paramètres avancés du modèle. Reportez-vous à la section Paramètres LLM avancés pour plus de détails sur la configuration d'Amazon Bedrock Guardrails, le débit provisionné pour Amazon Bedrock et les paramètres supplémentaires du modèle.

Inférence entre régions

L'inférence entre régions aide les utilisateurs d'Amazon Bedrock à gérer facilement les pics de trafic imprévus en utilisant le calcul dans différentes régions AWS. Pour utiliser l'inférence entre régions, vous avez besoin du profil d'inférence. Un profil d'inférence est une abstraction d'un pool de ressources à la demande provenant d'un ensemble configuré de régions AWS. Il peut acheminer votre demande d'inférence, provenant de votre région source, vers une autre région configurée dans ce pool. Cela permet de répartir le trafic entre plusieurs régions AWS. Cela permet d'augmenter le débit et d'améliorer la résilience pendant les périodes de pointe.

Les profils d'inférence sont nommés d'après le modèle et les régions qu'ils prennent en charge. Vous devez appeler un profil d'inférence depuis l'une des régions qu'il inclut. Par exemple, comme indiqué dans le tableau suivant, l'ID du profil d'inférence `us.anthropic.claude-3-haiku-20240307-v1:0` permet de répartir le trafic sur `us-east-1` et `us-west-2` les régions du modèle que vous choisissez. Certains modèles ne sont disponibles qu'avec un profil d'inférence dans une région donnée.

Profil d’inférence	ID du profil d'inférence	Régions incluses
Haïku américain Anthropic Claude 3	`us.anthropic.claude-3-haiku-20240307-v1:0`	USA Est (Virginie du Nord) (`us-east-1`) USA Ouest (Oregon) (`us-west-2`)

Si vous souhaitez utiliser un ID de profil d'inférence au lieu d'un ID de modèle, vous devez identifier l'ID de profil d'inférence approprié. Consultez la section Régions et modèles pris en charge pour les profils d'inférence dans le guide de l'utilisateur d'Amazon Bedrock pour plus d'informations. Dans la

console Amazon Bedrock, l'option d'inférence entre régions dans le menu de navigation de gauche fournit ces profils d'inférence. IDs

Après avoir identifié l'ID du profil d'inférence à utiliser, vous pouvez l'utiliser lors de l'étape de sélection du modèle en effectuant les étapes suivantes :

1. Sélectionnez Amazon Bedrock comme fournisseur de modèles.
2. Sélectionnez l'option du bouton radio Inference Profiles.
3. Entrez l'ID de votre profil d'inférence dans la zone de texte qui apparaît.

Reportez-vous à la section Améliorez la résilience grâce à l'inférence interrégionale dans le guide de l'utilisateur d'Amazon Bedrock pour plus de détails sur les profils d'inférence.

Sélectionnez la base de connaissances

Si vous souhaitez déployer un cas d'utilisation de génération augmentée sans extraction (RAG), vous pouvez ignorer cette étape.

Toutefois, si vous souhaitez activer RAG dans le cadre de votre déploiement, vous pouvez désormais fournir un identifiant d'index Amazon Kendra préconfiguré ou un identifiant de base de connaissances Amazon Bedrock. Vous pouvez également créer un nouvel index Amazon Kendra à utiliser avec la solution. La solution prend actuellement en charge les bases de connaissances Amazon Kendra et Amazon Bedrock en tant que bases de connaissances pour votre déploiement de cas d'utilisation basé sur RAG.

Reportez-vous à la section Configuration d'une base de connaissances pour obtenir des instructions sur l'ingestion de données dans la base de connaissances à utiliser avec votre déploiement basé sur RAG.

Configurations RAG avancées

L'assistant vous permet de sélectionner des options avancées à utiliser avec votre déploiement RAG, telles que le nombre de documents à récupérer chaque fois qu'une requête est envoyée à votre base de connaissances, une réponse texte statique du LLM lorsqu'aucun document n'est trouvé dans la base de connaissances, si vous souhaitez afficher les sources des documents avec votre réponse LLM pour les contrôles de santé, etc. Vous pouvez également configurer des configurations spécifiques à la base de connaissances pour Amazon Kendra, telles que le contrôle d'accès basé sur les rôles (RBAC) ou le type de recherche de remplacement lorsque vous utilisez Amazon Serverless

avec les bases de connaissances Amazon Bedrock OpenSearch . Reportez-vous à la section Paramètres avancés de la base de connaissances pour plus de détails sur ces paramètres avancés.

> **Note**
>
> Votre base de connaissances doit se trouver dans le même compte et dans la même région que le tableau de bord de déploiement déployé et les piles de cas d'utilisation.

Sélectionnez les invites et les limites de jetons

Au cours de cette étape, vous pouvez configurer votre invite pour une utilisation avec le LLM. Les invites peuvent nécessiter des espaces réservés tels que`{input}`, et`{history}`. `{context}` Ces espaces réservés indiquent au LLM où puiser les commentaires des utilisateurs, l'historique des conversations et les informations extraites de la base de connaissances.

- Pour le fournisseur de modèles Bedrock, l'invite du système doit être fournie, sans aucune restriction pour un cas d'utilisation autre que RAG. L'invite de désambiguïsation pour le fournisseur de modèles Bedrock nécessite toutefois un minimum de deux espaces réservés - et `{input}` `{history}`
- Pour le fournisseur de SageMaker modèles, le système et les instructions de désambiguïsation, les deux nécessitent un minimum de deux espaces réservés - et. `{input}` `{history}`
- Pour les cas d'utilisation de RAG, pour chaque fournisseur de modèles, l'`{context}`espace réservé est également requis.

Pour plus d'informations, consultez la section Configuration de vos invites. Vous pouvez également vous référer à la section Conseils pour gérer les limites de jetons des modèles lors de la sélection des tailles limites de jetons pour vos instructions.

Activer la saisie multimodale

Cette étape vous permet d'activer les fonctionnalités de saisie multimodales pour votre cas d'utilisation. Lorsque cette option est activée, les utilisateurs peuvent télécharger et envoyer des images et des documents en même temps que leurs requêtes textuelles.

Types de fichiers pris en charge et contraintes :

- Images : jusqu'à 20 images par message. La taille de chaque image ne doit pas dépasser 3,75 Mo et 8 000 pixels de hauteur et de largeur. Formats pris en charge : png, jpeg, gif, webp

- Documents : jusqu'à 5 documents par message. La taille de chaque document ne doit pas dépasser 4,5 Mo. Formats pris en charge : pdf, csv, doc, docx, xls, xlsx, html, txt, md

Comment utiliser la saisie multimodale :

1. Activer le MultimodalEnabledparamètre lors du déploiement des cas d'utilisation
2. Dans l'interface de chat, les utilisateurs peuvent télécharger des fichiers de deux manières :
   - En cliquant sur le bouton de téléchargement dans la zone de saisie du chat, ou
   - Glisser-déposer des fichiers directement dans l'interface de chat
3. Les fichiers sont chargés sur Amazon S3 et traités par le modèle sélectionné
4. Les fichiers téléchargés sont automatiquement supprimés au bout de 48 heures

Suivi de l'état des fichiers :

DevOps les utilisateurs peuvent surveiller les métadonnées des fichiers dans DynamoDB, notamment le temps de téléchargement et l'état du traitement. Les fichiers peuvent avoir les statuts suivants :

- en attente - Le téléchargement du fichier a été lancé mais n'est pas encore terminé. Il s'agit de l'état initial lorsqu'une URL présignée est générée.
- uploadé - Le fichier a été chargé avec succès sur S3 et est prêt à être traité par le modèle.
- supprimé - Le fichier a été supprimé par l'utilisateur et ne devrait plus être accessible pour le traitement.
- non valide - Les contrôles de validation du fichier ont échoué (par exemple, incompatibilité du type de fichier ou échec de la validation de sécurité).

Les fichiers en attente qui ne sont jamais téléchargés seront automatiquement nettoyés à l'expiration de leur TTL. Seuls les fichiers ayant le statut de téléchargement peuvent être traités par le modèle.

Le bucket multimodal S3 et la table de métadonnées DynamoDB sont disponibles dans les sorties du tableau de bord de déploiement avec les `MultimodalDataBucketName` clés et, respectivement. `MultimodalDataMetadataTable`

> **Note**
>
> Tous les modèles ne prennent pas en charge la saisie multimodale. Assurez-vous que le modèle sélectionné prend en charge le traitement des images et des documents avant

d'activer cette fonctionnalité. Reportez-vous aux modèles de base pris en charge dans la documentation d'Amazon Bedrock pour savoir quel modèle prend en charge l'image comme modalité d'entrée.

Important

Les fichiers chargés par les utilisateurs sont stockés dans Amazon S3 selon une politique de cycle de vie de 48 heures. Les métadonnées relatives aux fichiers chargés sont stockées dans Amazon DynamoDB avec un TTL de 24 heures pour l'historique des conversations.

Vérification et déploiement

Après cette étape, passez en revue les paramètres que vous avez sélectionnés et choisissez Deploy Use Case. Le nouveau cas d'utilisation est ensuite déployé et devient visible dans la vue du tableau de bord de déploiement afin de poursuivre la gestion.

Étape 3b : Déployer un cas d'utilisation de l'agent Bedrock

Le cas d'utilisation de Bedrock Agent fournit un mécanisme puissant et sécurisé pour invoquer des agents Amazon Bedrock dans le cadre de vos cas d'utilisation. Cette fonctionnalité permet aux développeurs d'intégrer de manière fluide les capacités des agents autonomes alimentés par l'IA, capables d'orchestrer et d'exécuter des tâches en plusieurs étapes sur différents modèles de base, sources de données, applications logicielles et conversations avec les utilisateurs, tout en maintenant des mesures de sécurité robustes.

Conditions préalables

Avant de créer un agent Amazon Bedrock, assurez-vous de disposer des éléments suivants :

1. Le compte AWS sur lequel Generative AI Application Builder sur AWS est déployé, avec un accès à la console Amazon Bedrock.
2. Autorisations IAM appropriées pour créer et gérer des agents Amazon Bedrock.

Création d'un agent Amazon Bedrock

Reportez-vous à la section Créer et configurer manuellement un agent dans le guide de l'utilisateur d'Amazon Bedrock pour obtenir des instructions détaillées sur la création d'un agent. Vous pouvez configurer des options telles que :

- Instructions (instructions) pour votre agent
- Base de connaissances, qui est utilisée pour rechercher des informations supplémentaires en fonction des entrées de l'utilisateur
- Mémoire de l'agent pour permettre aux agents de mémoriser des informations au cours de plusieurs sessions (pendant un maximum de 30 jours)

Une fois que vous avez créé avec succès un agent Amazon Bedrock, vous pouvez passer au flux de l'assistant de création d'applications Generative AI sur AWS Bedrock Agent. Pour ce faire, choisissez Déployer un nouveau cas d'utilisation sur le tableau de bord de déploiement et sélectionnez Créer un cas d'utilisation de l'agent Bedrock. Suivez l'assistant et suivez les étapes suivantes pour configurer le cas d'utilisation.

Sélectionnez un cas d'utilisation

Cette étape est identique au cas d'utilisation du texte décrit précédemment.

Sélectionnez la configuration réseau

Cette étape est identique au cas d'utilisation du texte décrit précédemment.

Sélectionnez un agent

Au cours de cette étape, vous devez fournir l'identifiant d'agent et l'identifiant d'alias de l'agent Amazon Bedrock que vous avez créé.

Étape 3c : Déployer un cas d'utilisation d'un serveur MCP

Le cas d'utilisation du serveur MCP (Model Context Protocol) vous permet de déployer et de gérer des serveurs MCP qui peuvent être intégrés à des modèles et agents d'IA. Les serveurs MCP fournissent un moyen standardisé d'exposer les outils, les ressources et les fonctionnalités aux applications d'intelligence artificielle. Vous pouvez soit créer des serveurs MCP à partir de fonctions Lambda existantes, soit héberger APIs des serveurs MCP personnalisés à l'aide d'images de conteneur.

Conditions préalables

Avant de déployer un cas d'utilisation du serveur MCP, assurez-vous que vous disposez des éléments suivants :

1. Le compte AWS sur lequel Generative AI Application Builder sur AWS est déployé.
2. Autorisations IAM appropriées pour créer et gérer les ressources Amazon Bedrock AgentCore .
3. En fonction de la méthode de création que vous avez choisie :
   - Pour la méthode Gateway (Lambda/API/MCPserveur) : fonctions Lambda, points de terminaison d'API avec leurs fichiers de schéma correspondants (format JSON pour Lambda, OpenAPI/ Smithy pour APIs) ou points de terminaison URL du serveur MCP
   - Pour la méthode d'exécution (ECR) : une image de conteneur Docker envoyée à Amazon ECR contenant l'implémentation de votre serveur MCP

Méthodes de création du serveur MCP

La solution prend en charge deux méthodes pour créer des serveurs MCP :

Création à partir d'un serveur Lambda, API ou MCP (méthode Gateway)

Cette méthode crée une passerelle MCP qui enveloppe les fonctions Lambda existantes, les serveurs REST ou les serveurs MCP externes APIs, les rendant accessibles en tant qu'outils MCP. La passerelle gère la traduction des protocoles entre MCP et vos services existants.

- Cibles Lambda : intégrez les fonctions Lambda existantes en fournissant l'ARN de la fonction et un fichier de schéma JSON décrivant le format de la fonction input/output
- Cibles OpenAPI : intégrez REST à l'aide des spécifications APIs OpenAPI (format JSON ou YAML) avec prise en charge de l'authentification 2.0 ou par clé API OAuth
- Cibles Smithy : intégration APIs définie à l'aide de fichiers de modèle Smithy (format .smithy ou .json)
- Cibles du serveur MCP : connectez-vous directement aux serveurs MCP externes via des points de terminaison URL, ce qui permet l'intégration de serveurs MCP existants sans déployer de nouvelle infrastructure

Vous pouvez configurer plusieurs cibles (jusqu'à 10) au sein d'une même passerelle MCP, chacune représentant un outil ou une fonctionnalité différent.

Hébergement à partir d'une image ECR (méthode d'exécution)

Cette méthode déploie un serveur MCP conteneurisé à partir d'une image Amazon ECR. Utilisez cette approche lorsque vous avez une implémentation de serveur MCP personnalisée qui doit être exécutée en tant que service autonome.

- Fournissez l'URI de l'image ECR (doit inclure une balise, par exemple, `:latest` ou `:v1.0.0`)
- Configurez éventuellement des variables d'environnement pour transmettre la configuration à votre conteneur
- Le conteneur doit implémenter le protocole MCP et exposer les points de terminaison requis

Déploiement d'un serveur MCP

Pour déployer un cas d'utilisation du serveur MCP, choisissez Déployer un nouveau cas d'utilisation sur le tableau de bord de déploiement, puis sélectionnez Créer un cas d'utilisation du serveur MCP. Suivez l'assistant et suivez les étapes suivantes pour configurer le cas d'utilisation.

Sélectionnez un cas d'utilisation

Cette étape est identique au cas d'utilisation du texte décrit précédemment.

Sélectionnez la configuration réseau

Actuellement, seul l'accès public est activé et le VPC n'est pas pris en charge pour la configuration réseau.

Création d'un serveur MCP

Au cours de cette étape, vous configurez le déploiement de votre serveur MCP :

Méthode de création du serveur MCP

Choisissez entre les deux méthodes de création :

- Création à partir d'un serveur Lambda, API ou MCP : créez une passerelle MCP à partir de fonctions Lambda existantes, de spécifications d'API ou de points de terminaison de serveur MCP externes
- Hébergement à partir d'une image ECR : Déployez un serveur MCP personnalisé à partir d'une image de conteneur

> **Note**
>
> La méthode de création ne peut pas être modifiée après le déploiement. Si vous devez changer de méthode, vous devez déployer un nouveau cas d'utilisation du serveur MCP.

Configuration de la passerelle (pour la méthode Lambda/API/MCP serveur)

Si vous avez sélectionné la méthode Gateway, configurez une ou plusieurs cibles :

1. Nom de la cible (obligatoire) : nom convivial pour identifier cette configuration cible
2. Description de la cible (facultatif) : brève description de ce que fait cette cible
3. Type de cible : sélectionnez le type de cible à configurer :
   - Lambda : pour les fonctions AWS Lambda
   - OpenAPI : pour REST avec les spécifications APIs OpenAPI
   - Smithy : Pour les définitions APIs du modèle Smithy
   - Serveur MCP : pour une connexion directe à des serveurs MCP externes via des points de terminaison URL
4. Fichier de schéma (obligatoire) : téléchargez le fichier de schéma qui décrit votre cible :
   - Pour Lambda : fichier de schéma JSON décrivant input/output le format. Pour plus d'informations sur la création de schémas d'outils Lambda, consultez le schéma d'outil Lambda dans le manuel Amazon Bedrock Developer Guide. AgentCore
   - Pour OpenAPI : fichier de spécification OpenAPI (JSON ou YAML). Pour en savoir plus sur les exigences du schéma OpenAPI, consultez le schéma OpenAPI dans le manuel Amazon Bedrock Developer Guide. AgentCore
   - Pour Smithy : fichier de modèle Smithy (.smithy ou .json). Pour en savoir plus sur la création de cibles Smithy, consultez la section Building Smithy targets dans le guide du développeur Amazon Bedrock AgentCore .
5. ARN de la fonction Lambda (obligatoire pour les cibles Lambda) : ARN de la fonction Lambda à intégrer
6. URL du serveur MCP (obligatoire pour les cibles du serveur MCP) : point de terminaison URL du serveur MCP externe auquel se connecter. L'URL doit être correctement encodée et le serveur MCP doit prendre en charge les fonctionnalités de l'outil avec les versions du protocole MCP 2025-06-18. Pour plus d'informations, consultez la section relative aux cibles des serveurs MCP dans le manuel Amazon Bedrock AgentCore Developer Guide.

7. Authentification sortante (obligatoire pour les cibles OpenAPI) : configurez l'authentification pour les appels d'API REST :
   - Type d'authentification : choisissez OAuth 2.0 ou clé API
   - ARN du fournisseur d'authentification sortant : ARN du fournisseur d'informations d'identification dans le coffre à jetons Amazon Bedrock AgentCore
   - Configurations supplémentaires : selon le type d'authentification :
     - Pour la OAuth version 2.0 : configurer les étendues et les paramètres personnalisés
     - Pour la clé d'API : spécifiez l'emplacement (en-tête ou paramètre de requête), le nom du paramètre et le préfixe facultatif

Vous pouvez ajouter plusieurs cibles (jusqu'à 10) en choisissant Ajouter une autre cible. Chaque cible représente un outil ou une fonctionnalité distinct exposé par votre serveur MCP.

Configuration ECR (pour la méthode ECR Image)

Si vous avez sélectionné la méthode Runtime, fournissez :

1. URI de l'image ECR (obligatoire) : l'URI complet de votre image Docker dans Amazon ECR
   - Format : `account-id.dkr.ecr.region.amazonaws.com/repository-name:tag`
   - L'image doit se trouver dans la même région AWS que votre déploiement
   - Une étiquette est requise (par exemple`:latest`,`:v1.0.0`)
2. Variables d'environnement (facultatives) : configurez les paires clé-valeur à transmettre à votre conteneur lors de l'exécution
   - Utilisez-les pour fournir une configuration, des informations d'identification ou des indicateurs personnalisés
   - Vous pouvez ajouter jusqu'à 10 variables d'environnement

Vérification et déploiement

Après avoir configuré votre serveur MCP, passez en revue les paramètres que vous avez sélectionnés et choisissez Deploy Use Case. Le nouveau cas d'utilisation du serveur MCP est ensuite déployé et devient visible dans la vue du tableau de bord de déploiement pour une gestion ultérieure.

> **Note**
>
> Les déploiements de serveurs MCP créent des ressources dans Amazon Bedrock AgentCore, notamment des passerelles, des environnements d'exécution et des identités de charge de travail. Ces ressources sont automatiquement gérées par la solution et seront nettoyées lorsque vous supprimerez le cas d'utilisation.

Étape 3d : Déployer un cas d'utilisation d'Agent Builder

L'Agent Builder vous permet de créer, de configurer et de déployer des agents d'IA prêts à être utilisés en production sur Amazon Bedrock. AgentCore Cette fonctionnalité permet de contrôler totalement le comportement des agents grâce aux instructions du système, à la sélection du modèle, à l'intégration du serveur MCP et à la gestion de la mémoire.

Le processus de déploiement est essentiellement le même que pour un cas d'utilisation en mode texte, avec quelques différences notables.

Sélectionnez un cas d'utilisation

Cette étape est identique au cas d'utilisation du texte décrit précédemment.

Détails du cas d’utilisation

Cette étape est identique au cas d'utilisation du texte décrit précédemment.

Configuration de l'agent

Au cours de cette étape, vous configurez les paramètres principaux de l'agent, notamment l'invite système, les servers/Strands outils MCP disponibles et la mémoire.

Prompt du système

L'invite du système définit le comportement, la personnalité et les capacités de l'agent. Vous pouvez effectuer les actions suivantes :

- Modifier le modèle d'invite système par défaut
- Utilisez le bouton Rétablir par défaut pour restaurer le modèle d'origine
- Incluez des instructions pour l'utilisation de l'outil et le formatage des réponses

Intégration au serveur MCP (facultatif)

Configurez les serveurs Model Context Protocol pour permettre à votre agent d'accéder aux outils et aux données de l'entreprise :

1. Sélectionnez l'un des serveurs MCP disponibles dans le menu déroulant
2. Passez en revue les outils prêts à l'emploi qui seront accessibles à l'agent

> **Note**
>
> Les serveurs MCP doivent être configurés et accessibles avant le déploiement. Reportez-vous à la documentation MCP pour les instructions de configuration du serveur.

Configuration de la mémoire

Configurez la manière dont l'agent gère le contexte et les connaissances :

- Mémoire à court terme : activée par défaut pour tous les agents. Maintient le contexte des conversations au cours des sessions.
- Mémoire à long terme : basculez pour activer l'extraction et le stockage des informations entre les sessions. Utilise AgentCore la mémoire avec une stratégie de mémoire sémantique.

Vérification et déploiement

Après cette étape, passez en revue les paramètres que vous avez sélectionnés et choisissez Deploy Use Case. Le déploiement d'Agent Builder prend généralement 10 à 15 minutes. Le nouveau cas d'utilisation devient alors visible dans la vue du tableau de bord de déploiement pour une gestion plus poussée.

Étape 3e : Déployer un cas d'utilisation du flux de travail

Le générateur de flux de travail vous permet de créer des agents superviseurs qui orchestrent plusieurs agents Agent Builder à l'aide du modèle de délégation Agents as Tools. Cette fonctionnalité vous permet de créer des flux de travail multi-agents complexes en réutilisant les déploiements d'Agent Builder existants.

Le processus de déploiement suit un schéma similaire à celui d'Agent Builder, avec des étapes supplémentaires pour la découverte et la sélection des agents.

Sélectionnez un cas d'utilisation

Cette étape est identique au cas d'utilisation du texte décrit précédemment.

Détails du cas d'utilisation

Cette étape est identique au cas d'utilisation du texte décrit précédemment.

Configuration de l'agent de supervision

Au cours de cette étape, vous configurez l'agent superviseur qui coordonnera les agents spécialisés Agent Builder.

Prompt du système

L'invite du système définit la manière dont l'agent superviseur délègue le travail aux agents spécialisés. Vous pouvez effectuer les actions suivantes :

- Modifier le modèle d'invite système par défaut
- Incluez des instructions pour la sélection et la délégation des agents
- Définissez comment agréger les résultats de plusieurs agents
- Utilisez le bouton Rétablir les paramètres par défaut pour restaurer le modèle d'origine

> **Note**
>
> L'invite du système doit clairement décrire quand et comment utiliser chaque agent spécialisé. Les descriptions des agents sont essentielles pour une délégation appropriée.

Sélection du modèle

Sélectionnez le modèle de base pour l'agent superviseur. L'agent superviseur utilise ce modèle pour :

- Comprendre les demandes des utilisateurs
- Sélectionnez les agents spécialisés appropriés
- Coordonner l'exécution des agents
- Agréger et mettre en forme les réponses

Sélectionnez des agents spécialisés

Au cours de cette étape, vous sélectionnez les agents Agent Builder auxquels le superviseur peut déléguer le travail.

Ajouter des agents

1. Cliquez sur Ajouter un agent pour ouvrir la boîte de dialogue de sélection de l'agent
2. Sélectionnez un ou plusieurs agents Agent Builder dans la liste
3. Passez en revue les descriptions des agents qui seront fournies au superviseur
4. Confirmez la sélection

> **Note**
>
> - Les flux de travail nécessitent au moins un cas d'utilisation d'Agent Builder en tant qu'agent spécialisé
> - Tous les agents spécialisés doivent être déployés avec succès avant de créer le flux de travail

Vérification et déploiement

Passez en revue la configuration du flux de travail, notamment :

- Prompt et modèle du système de l'agent superviseur
- Liste des agents spécialisés
- Réglages de mémoire

Choisissez Deploy Use Case. Le déploiement du flux de travail prend généralement 15 à 20 minutes. Le nouveau flux de travail devient visible dans la vue du tableau de bord de déploiement pour une gestion plus poussée.

Étape 4 : Configuration après le déploiement

Cette section fournit des recommandations pour configurer la solution après le déploiement.

Versionnage des compartiments Amazon S3, politiques de cycle de vie et réplication entre régions

Cette solution n'impose pas de configurations de cycle de vie aux compartiments qu'elle crée. Nous vous recommandons la procédure suivante :

- Définition des configurations du cycle de vie pour les déploiements de production. Pour plus de détails, consultez la section Configuration de la configuration du cycle de vie d'un bucket dans le guide de l'utilisateur d'Amazon Simple Storage Service.
- Activation du versionnement et de la réplication entre régions pour les compartiments Amazon S3 en fonction du cas d'utilisation pour lequel la solution est déployée.

Sauvegardes Amazon DynamoDB

Cette solution utilise DynamoDB à plusieurs fins (voir les services AWS dans cette solution). La solution n'active pas les sauvegardes pour les tables qu'elle crée. Nous recommandons de créer une sauvegarde de cette fonctionnalité pour les déploiements de production. Consultez Sauvegarde d'une table DynamoDB et Utilisation d'AWS Backup pour DynamoDB pour plus de détails.

CloudWatch Tableau de bord et alarmes Amazon

La solution déploie un tableau de bord personnalisé CloudWatch pour afficher des graphiques à partir de métriques publiées personnalisées et de métriques de service AWS. Nous vous recommandons de créer des CloudWatch alarmes et d'ajouter des notifications en fonction du cas d'utilisation pour lequel la solution est déployée.

Amazon CloudWatch Logs

Les journaux Lambda sont configurés pour ne jamais expirer et les journaux API Gateway sont configurés pour une expiration de 10 ans. Vous pouvez mettre à jour l'expiration des groupes de journaux respectifs afin de l'aligner sur la politique de conservation des enregistrements de votre entreprise.

Domaines Web personnalisés avec certificats TLS v1.2 ou supérieur

La solution déploie une interface utilisateur Web et une API Gateway optimisée pour Edge à l'aide CloudFront de. CloudFrontLe domaine n'applique pas les certificats TLS v1.2 ou supérieurs. Nous

vous recommandons de créer un domaine personnalisé à l'aide d'Amazon Route 53, de créer un certificat à l'aide d'AWS Certificate Manager ou d'utiliser un certificat existant si votre organisation en possède un.

Pour plus de détails, reportez-vous au guide du développeur Amazon Route 53 et à Choisir une version minimale de TLS pour un domaine personnalisé dans API Gateway.

Évoluer avec Amazon Kendra

Cette solution permet d'utiliser Amazon Kendra pour effectuer une recherche intelligente basée sur le langage naturel dans les documents ingérés. Vous pouvez augmenter la capacité d'Amazon Kendra en utilisant les CloudFormation paramètres suivants pour des charges de travail plus importantes :

Paramètre	Par défaut	Description
Capacité de requête supplémentaire d'Amazon Kendra	`0`	La quantité de capacité de requête supplémentaire pour un index et une GetQuerySuggestionscapacité. Une unité de capacité supplémentaire pour un index fournit environ 8 000 requêtes par jour.
Capacité de stockage supplémentaire d'Amazon Kendra	`0`	Quantité de capacité de stockage supplémentaire pour un index. Une unité de capacité unique fournit 30 Go d'espace de stockage ou 100 000 documents, selon la première éventualité.
Édition Amazon Kendra	`Developer`	Amazon Kendra propose des éditions Developer et Enterprise pour créer des index. Pour plus d'informations sur les différences entre les éditions Amazon Kendra, consultez les tarifs d'Amazon Kendra.

Pour modifier les valeurs de ces CloudFormation paramètres, sélectionnez les valeurs appropriées au moment du déploiement de la pile. Pour plus d'informations sur les unités de capacité de requête et de stockage, consultez la section Ajustement de la capacité.

> **Note**
>
> Si le cas d'utilisation du texte n'est pas déployé avec RAG activé, aucun index Amazon Kendra n'est utilisé ou créé.

Configuration du SSO à l'aide de la fédération Idp

Cette solution permet l'intégration avec des fournisseurs d'identité externes qui prennent en charge la fédération d'identité basée sur SAML ou OIDC. Lorsque la solution est déployée, elle crée un groupe d'utilisateurs Amazon Cognito et une intégration de clients d'applications individuels pour le tableau de bord de déploiement et les cas d'utilisation individuels. En fonction de l'Idp externe, suivez les étapes décrites dans la section Configuration des fournisseurs d'identité pour votre groupe d'utilisateurs du guide du développeur Amazon Cognito et choisissez l'intégration du client d'application pour le tableau de bord de déploiement ou le cas d'utilisation avec lequel vous souhaitez configurer l'authentification unique.

Pour transmettre les informations des groupes d'utilisateurs à la base de connaissances ou aux magasins vectoriels dans une architecture basée sur RAG, vous devez mapper les groupes d'utilisateurs de l'Idp externe aux groupes d'utilisateurs Amazon Cognito. La solution fournit un déclencheur initial de la fonction Lambda d'échafaudage à associer à la phase préalable à la génération du jeton. La fonction Lambda possède le fichier group_mapping.json qui doit être mis à jour pour fournir les mappages de groupes. Reportez-vous à la section Personnalisation des flux de travail des groupes d'utilisateurs avec des déclencheurs Lambda pour les déclencheurs Lambda pris en charge par Amazon Cognito.

Configuration manuelle du pool d'utilisateurs

Si vous choisissez de ne pas transmettre d'e-mail d'administrateur ou d'utilisateur par défaut lors du déploiement, vous devez créer manuellement les groupes d'utilisateurs appropriés dans Amazon Cognito afin de garantir les autorisations correctes :

1. Pour le tableau de bord de déploiement, créez un groupe nommé `Admin` dans votre groupe d'utilisateurs Cognito.

2. Pour chaque cas d'utilisation, créez un groupe nommé `${UseCaseName}-Users` dans votre groupe d'utilisateurs Cognito, où `${UseCaseName}` est le nom de votre cas d'utilisation déployé.

Ces groupes sont nécessaires au bon fonctionnement du mécanisme d'autorisation. Tous les utilisateurs auxquels vous souhaitez accorder l'accès doivent être ajoutés aux groupes appropriés.

En cas `placeholder@example.com` de réussite, le groupe Cognito sera créé, mais vous devez tout de même créer les utilisateurs associés et les affecter au groupe.

Personnalisation de l'écran de connexion

Cette solution utilise l'interface utilisateur hébergée par Amazon Cognito pour afficher la page de connexion. Pour personnaliser la page de connexion intégrée, reportez-vous à la section Personnalisation des pages Web de connexion et d'inscription intégrées dans le manuel Amazon Cognito Developer Guide.

Considérations supplémentaires en matière de sécurité

En fonction du cas d'utilisation pour lequel vous déployez la solution, consultez les recommandations de sécurité suivantes :

- Clés de chiffrement AWS KMS gérées par le client : la solution utilise les clés AWS KMS gérées par AWS par défaut, car celles-ci sont disponibles sans frais supplémentaires. Passez en revue votre cas d'utilisation pour déterminer si vous devez mettre à jour la solution pour utiliser les clés AWS KMS gérées par le client.
- Règles de régulation d'API Gateway : la solution est déployée avec des règles de régulation par défaut sur API Gateway. En fonction de votre cas d'utilisation et des volumes de transactions attendus, nous vous recommandons de configurer la régulation pour le. APIs Pour plus de détails, consultez les demandes d'API Throttle pour un meilleur débit dans le manuel Amazon API Gateway Developer Guide.
- Activation d'AWS CloudTrail : comme pratique de sécurité recommandée, pensez à activer AWS CloudTrail dans le compte AWS sur lequel la solution est déployée pour consigner les appels d'API dans le compte AWS. Pour plus de détails, consultez le guide de CloudTrail l'utilisateur AWS.
- Détection de la dérive - Nous recommandons de configurer la détection de la dérive sur les CloudFormation piles afin d'identifier et d'être averti des modifications involontaires ou malveillantes apportées à la pile de solutions déployée. Pour plus de détails, consultez Implémentation d'une alarme pour détecter automatiquement la dérive dans les CloudFormation piles AWS.

- Jetons Web Cognito JSON (JWTs) : la solution utilise Amazon Cognito pour s'authentifier auprès des points de terminaison JWTs de l'API REST. Nous avons configuré la solution avec une expiration de cinq minutes pour les jetons d'identification et les jetons d'accès. Lorsqu'un utilisateur se déconnecte, sa capacité à générer de nouveaux jetons est révoquée (le jeton d'actualisation est révoqué). Cependant, jusqu'à l'expiration du jeton actuel, toutes les demandes adressées au point de terminaison de l'API seront authentifiées avec succès, car elles disposent d'un jeton valide. Passez en revue les considérations de sécurité relatives à votre cas d'utilisation et ajustez la période de validité du jeton.

Personnalisation des politiques de cycle de vie :

Pour les déploiements en production, passez en revue et ajustez les politiques de cycle de vie en fonction de vos exigences de rétention. Consultez la section Configuration du cycle de vie d'un bucket dans le guide de l'utilisateur d'Amazon Simple Storage Service.

Stockage et cycle de vie des fichiers multimodaux

Si vous avez activé les fonctionnalités de saisie multimodales (MultimodalEnableddéfinies sur`Yes`) pour votre cas d'utilisation, la solution crée un compartiment Amazon S3 pour stocker les fichiers téléchargés et une table DynamoDB pour suivre les métadonnées des fichiers.

Politiques de cycle de vie par défaut :

- Fichiers S3 : supprimés automatiquement après 48 heures
- Métadonnées DynamoDB : les enregistrements expirent au bout de 24 heures (historique des conversations TTL)

Considérations relatives à la sécurité :

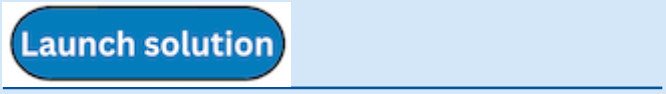
- Les fichiers sont partitionnés par ID de cas d'utilisation, ID utilisateur, ID de conversation et ID de message et un fichier est stocké avec un nom UUID à la place. Le mappage entre l'UUID et les noms de fichiers est disponible dans la table de métadonnées DynamoDB
- Les utilisateurs ne peuvent accéder qu'aux fichiers qu'ils ont téléchargés dans le cadre de leurs propres conversations
- La validation du type de fichier est effectuée à l'aide de la détection de nombres magiques
- Nous vous recommandons d'activer Amazon GuardDuty Malware Protection pour S3 afin de détecter tout contenu malveillant dans les fichiers téléchargés.

Déploiement d'un cas d'utilisation de texte autonome

Suivez les step-by-step instructions de cette section pour configurer et déployer la solution dans votre compte.

Temps de déploiement : environ 10 à 30 minutes

1. Connectez-vous à l'AWS Management Console et sélectionnez le bouton pour lancer le CloudFront modèle que vous souhaitez déployer.

BedrockChat.modèle	
SageMakerChat.modèle	

2. Le modèle est lancé par défaut dans la région USA Est (Virginie du Nord). Pour lancer la solution dans une autre région AWS, utilisez le sélecteur de région dans la barre de navigation de la console.

   Remarque : Cette solution utilise Amazon Kendra et Amazon Bedrock, qui ne sont actuellement pas disponibles dans toutes les régions AWS. Si vous utilisez ces fonctionnalités, vous devez lancer cette solution dans une région AWS où ces services sont disponibles. Pour connaître la disponibilité la plus récente par région, consultez la liste des services régionaux AWS.
3. Sur la page Create stack *, vérifiez que l'URL du modèle correcte se trouve dans la zone de texte *Amazon S3 URL *et choisissez *Next.
4. Sur la page *Spécifier les détails de la pile *, attribuez un nom à votre pile de solutions. Pour plus d'informations sur les limites relatives aux caractères de dénomination, consultez les limites IAM et STS dans le guide de l'utilisateur d'AWS Identity and Access Management.
5. Sous Paramètres, passez en revue les paramètres de ce modèle de solution et modifiez-les si nécessaire. Cette solution utilise les valeurs par défaut suivantes.

UseCaseUUID	*<_Requires input_>*	36 caractères UUIDv4 pour identifier ce cas d'utilisa tion déployé au sein d'une application.

UseCaseConfigRecordKey	*<_Requires input_>*	Clé correspondant à l'enregistrement contenant les configurations requises par le fournisseur de chat Lambda lors de l'exécution. L'enregistrement de la table doit avoir un attribut clé correspondant à cette valeur et un attribut de configuration contenant la configuration souhaitée. Cet enregistrement sera renseigné par la plateforme de déploiement s'il est utilisé. Pour les déploiements autonomes de ce cas d'utilisation, une entrée créée manuellement dans le tableau défini dans UseCaseConfigTableNameest requise.
UseCaseConfigTableName	*<_Requires input_>*	La pile lira la configuration à partir de la table avec ce nom comme clé UseCaseConfigRecordKey

ExistingRestApiId	(Entrée facultative)	ID d'API REST API Gateway existant à utiliser. Si elle n'est pas fournie, une nouvelle API REST API Gateway sera créée. Généralement fourni lors du déploiement depuis le tableau de bord de déploiement. Remarque : L'utilisation d'APIs Existing peut contribuer à réduire la duplication des ressources et à simplifier la gestion APIs lorsque vous devez déployer plusieurs cas d'utilisation autonomes. Lorsque vous fournissez APIs des données existantes pour un cas d'utilisation autonome, vous devez vous assurer que l'API est configurée avec les routes requises avec les modèles attendus. Une route /details préconfigurée obligatoire (récupère les détails des cas d'utilisation pendant le chat) et éventuellement une route /feedback (si elle FeedbackEnabledest définie pour permettre la collecte de commentaires pour les réponses de chat LLM) doivent `Yes` être configurées. En outre ExistingApiRootResourceId, ExistingCognitoUse

		rPoolIdet ExistingCognitoGroupPolicyTableNamedoit également être fourni.
ExistingApiRootResourceId	(Entrée facultative)	ID de ressource racine de l'API REST API Gateway existant à utiliser. L'ID de ressource racine de l'API REST peut être obtenu depuis la console AWS en sélectionnant la ressource racine (/) dans la section « Ressources » de l'API. L'ID de ressource sera ensuite affiché dans le panneau des détails de la ressource. Vous pouvez également exécuter un appel d'API de description sur votre API REST pour trouver l'ID de ressource racine.
FeedbackEnabled	No	Si ce paramètre est défini sur Non, la pile de cas d'utilisation déployée n'aura pas accès à la fonctionnalité de feedback.

ExistingModelInfoTableName	(Entrée facultative)	Nom de la table DynamoDB pour la table contenant les informations sur le modèle et les valeurs par défaut. Utilisé par la plateforme de déploiement. En cas d'omission, une nouvelle table sera créée pour héberger les valeurs par défaut du modèle.
DefaultUserEmail	`placeholder@example.com`	Adresse e-mail de l'utilisateur par défaut pour ce cas d'utilisation. Un utilisateur Amazon Cognito pour cet e-mail est créé pour accéder au cas d'utilisation. S'ils ne sont pas fournis, le groupe Cognito et l'utilisateur ne seront pas créés. Vous pouvez également l'utiliser `placeholder@example.com` pour créer le groupe, mais pas l'utilisateur. Reportez-vous à la section Configuration manuelle du groupe d'utilisateurs pour plus d'informations sur la configuration de votre groupe d'utilisateurs.

ExistingCognitoUserPoolId	(Entrée facultative)	UserPoolIdd'un groupe d'utilisateurs Amazon Cognito existant auprès duquel ce cas d'utilisation sera authentif ié. Généralement fourni lors du déploiement depuis le tableau de bord de déploieme nt, mais peut être omis lors du déploiement de cette pile de cas d'utilisation autonome.
CognitoDomainPrefix	(Entrée facultative)	Entrez une valeur si vous souhaitez fournir un domaine au client du pool d'utilisateurs Cognito. Si vous ne fournisse z aucune valeur, le déploieme nt en générera une.
ExistingCognitoUserPoolClie nt	(Entrée facultative)	Fournissez un client de groupe d'utilisateurs (client d'application) pour utiliser un client existant. Si vous ne fournissez pas de client de groupe d'utilisateurs, un nouveau client sera créé. Ce paramètre ne peut être fourni que si un identifiant de groupe d'utilisateurs existant est fourni.

ExistingCognitoGroupPolicyTableName	(Entrée facultative)	Nom de la table DynamoDB contenant les politiques de groupe d'utilisateurs. Ceci est utilisé par l'autorisateur personnalisé sur l'API du cas d'utilisation. En règle générale, vous pouvez fournir une entrée lors du déploiement depuis la plate-forme de déploiement, mais vous pouvez l'omettre lors du déploiement de cette pile de cas d'utilisation autonome.
RAGEnabled	`true`	Si ce paramètre est défini sur true, la pile de cas d'utilisation déployée utilise l'index Amazon Kendra fourni, créé pour fournir les fonctionnalités RAG. Si ce paramètre est défini sur`false`, l'utilisateur interagit directement avec le LLM.
KnowledgeBaseType	Bedrock	Type de base de connaissances à utiliser pour RAG. Ne définissez que si RAGEnabled c'est le cas`true`. Cela peut être Bedrock ou Kendra. Remarque : pertinent uniquement si RAGEnabled c'est vrai.

ExistingKendraIndexId	(Entrée facultative)	ID d'index d'un index Kendra existant à utiliser pour le cas d'utilisation. Si aucun index n'est fourni et Knowledge BaseTypequ'il s'agit de Kendra, un nouvel index sera créé pour vous. Remarque : N'est pertinent que s'il RAGEnabledest `true` et s'KnowledgeBaseTypeil est`Kendra`.
NewKendraIndexName	(Entrée facultative)	Nom du nouvel index Kendra à créer pour ce cas d'utilisation. Ne s'applique que s'ExistingKendraIndexIdil n'est pas fourni. Remarque : Uniquement pertinent si RAGEnabledc'est vrai et si c'KnowledgeBaseTypeest le cas de Kendra.

NewKendraQueryCapacityUnits	0	Des unités de capacité de requête supplémentaires pour le nouvel index Amazon Kendra seront créées pour ce cas d'utilisation. Ne s'applique que s'ExistingKendraIndexIdil n'est pas fourni, voir CapacityUnitsConfiguration. Remarque : N'est pertinent que s'il RAGEnabledest `true` et s'KnowledgeBaseTypeil est`Kendra`.
NewKendraStorageCapacityUnits	`0`	Des unités de capacité de stockage supplémentaires pour le nouvel index Amazon Kendra seront créées pour ce cas d'utilisation. Ne s'applique que s'ExistingKendraIndexIdil n'est pas fourni, voir CapacityUnitsConfiguration. Remarque : N'est pertinent que s'il RAGEnabledest `true` et s'KnowledgeBaseTypeil est`Kendra`.

NewKendraIndexEdition	(Entrée facultative)	L'édition d'Amazon Kendra à utiliser pour le nouvel index Amazon Kendra qui sera créé pour ce cas d'utilisation. S'applique uniquement s'il n'ExistingKendraIndexIdest pas fourni, voir Amazon Kendra Editions. Remarque : N'est pertinent que s'il RAGEnabledest `true` et s'KnowledgeBaseTypeil est`Kendra`.
BedrockKnowledgeBaseId	(Entrée facultative)	Identifiant de la base de connaissances sur le socle à utiliser dans un cas d'utilisation de RAG. Ne peut pas être fourni si ExistingKendraIndexIdou NewKendraIndexNamesont fournis. Remarque : N'est pertinent que s'il RAGEnabledest `true` et s'KnowledgeBaseTypeil est`Bedrock`.
VpcEnabled	No	Les ressources de la pile doivent-elles être déployées au sein d'un VPC ?

CreateNewVpc	No	SélectionnezYes, si vous souhaitez que la solution crée un nouveau VPC pour vous et qu'elle soit utilisée pour ce cas d'utilisation. Remarque : pertinent uniquement si tel VpcEnable dest le casYes.
IPAMPoolId	(Entrée facultative)	Si vous souhaitez attribuer la plage d'adresses CIDR à l'aide du gestionnaire d'adresses IP Amazon VPC, fournissez l'ID du pool IPAM à utiliser. Remarque : N'est pertinent que s'il VpcEnabledest Yes et s'CreateNewVpcil estNo.
ExistingVpcId	(Entrée facultative)	ID VPC d'un VPC existant à utiliser pour le cas d'utilisa tion. Remarque : N'est pertinent que s'il VpcEnabledest Yes et s'CreateNewVpcil estNo.
ExistingPrivateSubnetIds	(Entrée facultative)	Liste séparée par des virgules IDs des sous-rése aux privés existants à utiliser pour déployer la fonction Lambda. Remarque : N'est pertinent que s'il VpcEnabledest Yes et s'CreateNewVpcil estNo.

ExistingSecurityGroupIds	(Entrée facultative)	Liste séparée par des virgules des groupes de sécurité du VPC existant à utiliser pour configurer les fonctions Lambda. Remarque : N'est pertinent que s'il VpcEnabledest `Yes` et s'CreateNewVpcil est`No`.
VpcAzs	(Entrée facultative)	Liste séparée par des AZs virgules indiquant dans laquelle les sous-réseaux du VPCs sont créés Remarque : N'est pertinent que s'il VpcEnabledest `Yes` et s'CreateNewVpcil est`No`.
UseInferenceProfile	`No`	Si le modèle configuré est Bedrock, vous pouvez indiquer si vous utilisez le profil d'inférence Bedrock. Cela garantira que les politiques IAM requises seront configurées lors du déploiement de la pile. Pour plus de détails, reportez-vous au https://docs.aws.amazon.com/bedrock/ latest/userguide/cross fichier -region-inference.html suivant

Interface utilisateur de déploiement	Oui	Sélectionnez l'option permettant de déployer l'interface utilisateur frontale pour ce déploiement. Si vous sélectionnez Non, vous ne créerez que l'infrastructure pour héberger le APIs traitement principal APIs, l'authentification et le traitement principal.

6. Choisissez Next (Suivant).
7. Sur la page Configurer les options de pile, choisissez Suivant.
8. Sur la page Vérification, vérifiez et confirmez les paramètres. Cochez la case indiquant que le modèle créera des ressources AWS Identity and Access Management (IAM).
9. Sélectionnez Create stack (Créer une pile) pour déployer la pile.

Vous pouvez consulter l'état de la pile dans la CloudFormation console AWS dans la colonne Status. Vous devriez recevoir le statut CREATE_COMPLETE dans 10 à 30 minutes environ.

Déploiement d'un cas d'utilisation autonome de l'agent Bedrock

Suivez les step-by-step instructions de cette section pour configurer et déployer la solution dans votre compte.

Temps de déploiement : environ 10 à 30 minutes

1. Connectez-vous à l'AWS Management Console et sélectionnez le bouton pour lancer le CloudFront modèle.

BedrockAgent.modèle	Launch solution

2. Le modèle est lancé par défaut dans la région USA Est (Virginie du Nord). Pour lancer la solution dans une autre région AWS, utilisez le sélecteur de région dans la barre de navigation de la console.

> **Note**
>
> Cette solution utilise Amazon Bedrock, qui n'est actuellement pas disponible dans toutes les régions AWS. Si vous utilisez ces fonctionnalités, vous devez lancer cette solution dans une région AWS où ces services sont disponibles. Pour connaître la disponibilité la plus récente par région, consultez la liste des services régionaux AWS.

3. Sur la page Create stack, vérifiez que l'URL du modèle est correcte dans la zone de texte URL Amazon S3 et choisissez Next.
4. Sur la page Spécifier les détails de la pile, attribuez un nom à votre pile de solutions. Pour plus d'informations sur les limites relatives aux caractères de dénomination, consultez {https---docs-aws-amazon-com- https---docs-aws-amazon-com -IAM-latest- UserGuide -reference-iam-limits-html} [quotas IAM et AWS STS] dans le guide de l'utilisateur d'AWS Identity and Access Management.
5. Sous Paramètres, passez en revue les paramètres de ce modèle de solution et modifiez-les si nécessaire. Cette solution utilise les valeurs par défaut suivantes.

Paramètre	Entrée par défaut	Description
UseCaseUUID	*<_Requires input_>*	36 caractères UUIDv4 pour identifier ce cas d'utilisation déployé au sein d'une application.
UseCaseConfigRecordKey	*<Requires input>*	Clé correspondant à l'enregistrement contenant les configurations requises par la fonction Lambda du fournisseur de chat lors de l'exécution. L'enregistrement de la table doit avoir un attribut clé correspondant à cette valeur et un attribut de configuration contenant la configuration souhaitée.

Paramètre	Entrée par défaut	Description
		Cet enregistrement sera renseigné par la plateform e de déploiement si elle est utilisée. Pour les déploieme nts autonomes de ce cas d'utilisation, une entrée créée manuellement dans le tableau défini dans UseCaseConfigTable Nameest requise.
UseCaseConfigTableName	`<Requires input>`	La pile lira la configuration des cas d'utilisation à partir du tableau fourni ici et à l'aide de la clé d'enregistrement définie dans UseCaseCo nfigRecordKey.
DefaultUserEmail	`placeholder@exampl e.com`	Adresse e-mail de l'utilisa teur par défaut pour ce cas d'utilisation. La solution crée un utilisateur Amazon Cognito pour cet e-mail afin d'accéder au cas d'utilisation.

Paramètre	Entrée par défaut	Description
ExistingRestApiId	(Entrée facultative)	ID d'API REST API Gateway existant à utiliser. Si elle n'est pas fournie, une nouvelle API REST API Gateway sera créée. Généralement fourni lors du déploiement depuis le tableau de bord de déploiement. Remarque : L'utilisation d'APIs Existing peut contribuer à réduire la duplication des ressources et à simplifier la gestion APIs lorsque vous devez déployer plusieurs cas d'utilisation autonomes. Lorsque vous fournissez APIs des données existantes pour un cas d'utilisation autonome, vous devez vous assurer que l'API est configurée avec les routes requises avec les modèles attendus. Une route /details préconfigurée obligatoire (récupère les détails des cas d'utilisation pendant le chat) et éventuellement une route /feedback (si elle FeedbackEnabledest définie pour permettre la collecte de commentaires pour les réponses de chat LLM) doivent `Yes` être configurées. En outre ExistingApiRootRes

Paramètre	Entrée par défaut	Description
		ourceId, ExistingCognitoUse rPoolIdet ExistingCognitoGro upPolicyTableNamedoit également être fourni.
ExistingApiRootResourceId	(Entrée facultative)	ID de ressource racine de l'API REST API Gateway existant à utiliser. L'ID de ressource racine de l'API REST peut être obtenu à partir de la console AWS en sélectionnant la ressource racine (/) dans la section « Ressources » de l'API. L'ID de ressource sera ensuite affiché dans le panneau des détails de la ressource. Vous pouvez également exécuter un appel d'API de descripti on sur votre API REST pour trouver l'ID de ressource racine.
FeedbackEnabled	No	Si ce paramètre est défini sur Non, la pile de cas d'utilisa tion déployée n'aura pas accès à la fonctionnalité de feedback.

Paramètre	Entrée par défaut	Description
CognitoDomainPrefix	(Entrée facultative)	Entrez une valeur si vous souhaitez fournir un domaine au client du groupe d'utilisa teurs Amazon Cognito. Si vous ne fournissez aucune valeur, la solution en génère une.
ExistingCognitoUserPoolId	(Entrée facultative)	UserPoolIdd'un groupe d'utilisateurs Amazon Cognito existant auprès duquel vous souhaitez authentifier ce cas d'utilisation. REMARQUE : vous fournissez généralem ent cet identifiant lors du déploiement depuis le tableau de bord de déploiement, mais vous pouvez l'omettre lors du déploiement de cette pile de cas d'utilisation autonome.
ExistingCognitoUserPoolClie nt	(Entrée facultative)	Fournissez un client de groupe d'utilisateurs (client d'application) pour utiliser un client existant. Si vous ne fournissez pas de client de groupe d'utilisateurs, la solution en crée un. Vous ne pouvez fournir ce paramètre que si vous avez fourni un ExistingCognitoUserPoolId.

Paramètre	Entrée par défaut	Description
ExistingCognitoGroupPolicyTableName	(Entrée facultative)	Nom de la table DynamoDB contenant les politiques de groupe d'utilisateurs. Ceci est utilisé par l'autorisateur personnalisé sur l'API du cas d'utilisation. REMARQUE : vous fournissez généralement ce nom lors du déploiement depuis le tableau de bord de déploiement, mais vous pouvez l'omettre lors du déploiement de cette pile de cas d'utilisation autonome.
VpcEnabled	`No`	Si les ressources de la pile doivent être déployées au sein d'un VPC.
CreateNewVpc	`No`	Indiquez `Yes` si vous souhaitez que la solution crée un nouveau VPC pour vous et l'utilise dans ce cas d'utilisation. REMARQUE : Ce paramètre n'est pertinent que s'il l'VpcEnabledest`Yes`.
IPAMPoolId	(Entrée facultative)	Si vous souhaitez attribuer la plage d'adresses CIDR à l'aide d'IPAM, indiquez l'ID du pool IPAM à utiliser. REMARQUE : Ce paramètre n'est pertinent que s'il VpcEnabledest `Yes` et s'CreateNewVpcil est`No`.

Paramètre	Entrée par défaut	Description
ExistingVpcId	(Entrée facultative)	ID VPC d'un VPC existant à utiliser pour le cas d'utilisa tion. REMARQUE : Ce paramètre n'est pertinent que s'il VpcEnabledest `Yes` et s'CreateNewVpcil est`No`.
ExistingPrivateSubnetIds	(Entrée facultative)	Liste séparée par des virgules IDs des sous-rése aux privés existants à utiliser pour déployer la fonction Lambda. REMARQUE : Ce paramètre n'est pertinent que s'il VpcEnabledest `Yes` et s'CreateNewVpcil est`No`.
ExistingSecurityGroupIds	(Entrée facultative)	Liste séparée par des virgules des groupes de sécurité du VPC existant à utiliser pour configure r les fonctions Lambda. REMARQUE : Ce paramètre n'est pertinent que s'il VpcEnabledest `Yes` et s'CreateNewVpcil est`No`.
VpcAzs	(Entrée facultative)	Liste séparée par des AZs virgules indiquant dans laquelle les sous-réseaux du VPCs sont créés Remarque : N'est pertinent que s'il VpcEnabledest `Yes` et s'CreateNewVpcil est`No`.

Paramètre	Entrée par défaut	Description
BedrockAgentId	*<Requires input>*	L'ID de l'agent Amazon Bedrock à utiliser.
BedrockAgentAliasId	*<Requires input>*	L'identifiant d'alias de l'agent Amazon Bedrock à utiliser.
Interface utilisateur de déploiement	Yes	Sélectionnez l'option permettant de déployer l'interface utilisateur de discussion frontale pour ce déploiement. La sélection No entraîne la création de l'infrastructure pour héberger le APIs, l'authentification pour le APIs et le traitemen t principal sans l'interface utilisateur de chat.

6. Choisissez Next (Suivant).
7. Sur la page Configurer les options de pile, choisissez Suivant.
8. Sur la page Vérification, vérifiez et confirmez les paramètres. Cochez la case indiquant que le modèle créera des ressources IAM.
9. Sélectionnez Create stack (Créer une pile) pour déployer la pile.

Vous pouvez consulter l'état de la pile dans la CloudFormation console AWS dans la colonne Status. Vous devriez recevoir le statut CREATE_COMPLETE dans 10 à 30 minutes environ.

Fourniture d'une configuration de chat DynamoDB

Lors du déploiement d'un cas d'utilisation, UseCaseConfigRecordKeyUseCaseConfigTableNameles CloudFormation paramètres requis sont normalement renseignés par le tableau de bord de déploiement. La pile des tableaux de bord de déploiement gère la création et la configuration de cette table, tandis que les appels à l'API de déploiement déclenchent le peuplement des paramètres.

Lorsque vous effectuez un déploiement autonome, vous devez effectuer les opérations suivantes :

1. Créez une table DynamoDB avec une clé de hachage.
2. Créez un enregistrement dans le tableau contenant la configuration pour le cas d'utilisation sous forme d'enregistrement au format : `{key: some_use_case_key, config: {your_configuration}.`
3. Transmettez les paramètres choisis UseCaseConfigTableNameet UseCaseConfigRecordKey(`some_use_case_key`dans cet exemple) à la pile de cas d'utilisation lors du déploiement.

Pour créer une configuration adaptée à un déploiement autonome, vous pouvez créer un cas d'utilisation requis à partir du tableau de bord de déploiement et copier l'enregistrement depuis le tableau de configuration. Sinon, vous pouvez créer votre propre configuration en vous basant sur l'exemple suivant pour un déploiement de Bedrock :

```
{
 "UseCaseName": "SampleUseCase",
 "ConversationMemoryParams": {
 "ConversationMemoryType": "DynamoDB",
 "HumanPrefix": "H",
 "AiPrefix": "A",
 "ChatHistoryLength": 20
 },
 "KnowledgeBaseParams": {
 "KnowledgeBaseType": "Bedrock",
 "NumberOfDocs": 2,
 "ScoreThreshold": 0,
 "ReturnSourceDocs": false,
 "BedrockKnowledgeBaseParams": {
 "BedrockKnowledgeBaseId": "SOME_ID",
 "OverrideSearchType": null
 }
 },
 "LlmParams": {
 "ModelProvider": "Bedrock",
 "BedrockLlmParams": { "ModelId": "anthropic.claude-v2" },
 "PromptParams": {
 "PromptTemplate": "some prompt",
 "MaxPromptTemplateLength": 187500,
 "MaxInputTextLength": 187500,
 "UserPromptEditingEnabled": true,
 "DisambiguationEnabled": true,
```

```
 "DisambiguationPromptTemplate": "some prompt"
 },
 "ModelParams": {},
 "Temperature": 1,
 "RAGEnabled": true,
 "Streaming": true,
 "Verbose": false
 }
}
```

Surveillez la solution avec Service Catalog AppRegistry

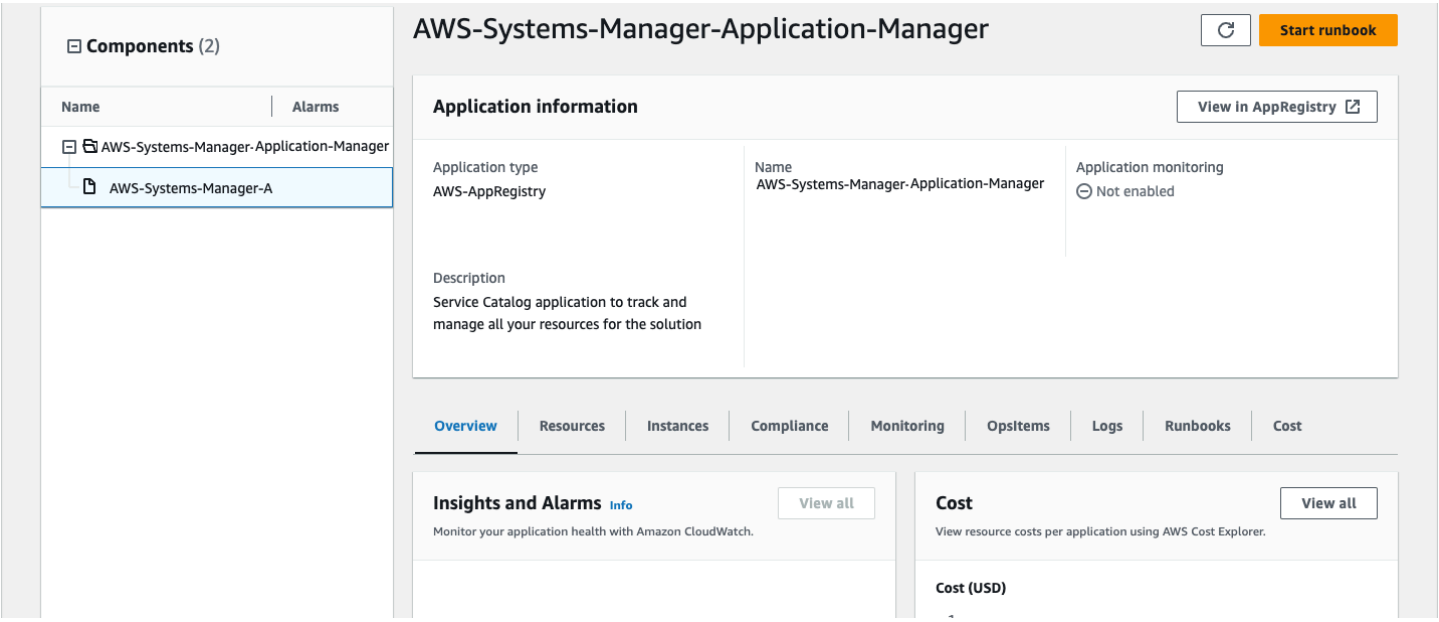
La solution inclut une AppRegistry ressource Service Catalog pour enregistrer le CloudFormation modèle et les ressources sous-jacentes en tant qu'application dans Service Catalog AppRegistry et Systems Manager Application Manager.

Systems Manager Application Manager vous donne une vue d'ensemble de cette solution et de ses ressources au niveau de l'application, afin que vous puissiez :

- Surveillez ses ressources, les coûts des ressources déployées sur les stacks et les comptes AWS, ainsi que les journaux associés à cette solution depuis un emplacement central.
- Affichez les données d'exploitation des ressources de cette solution dans le contexte d'une application. Par exemple, l'état du déploiement, les CloudWatch alarmes, les configurations des ressources et les problèmes opérationnels.

La figure suivante illustre un exemple de vue d'application pour la pile de solutions dans Application Manager.

Représente la pile de solutions dans le gestionnaire d'applications



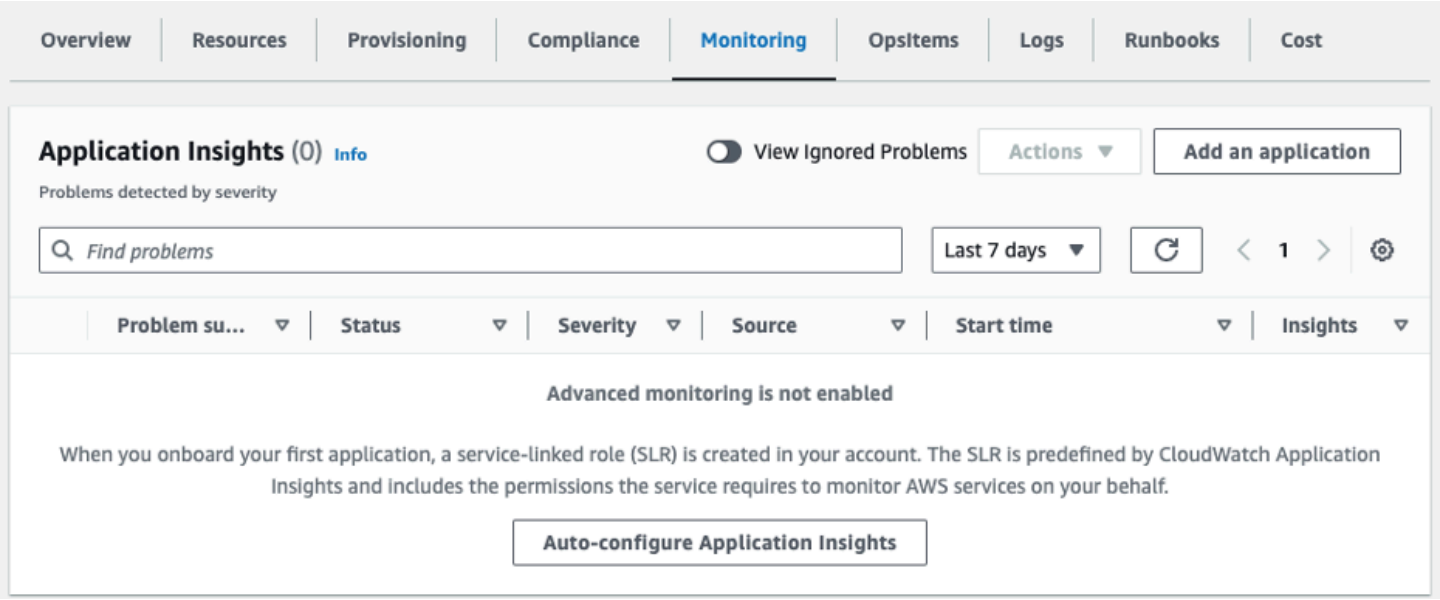
Activer CloudWatch Application Insights

1. Connectez-vous à la console Systems Manager.

2. Dans le volet de navigation, choisissez Application Manager.
3. Dans Applications, recherchez le nom de l'application pour cette solution et sélectionnez-la.

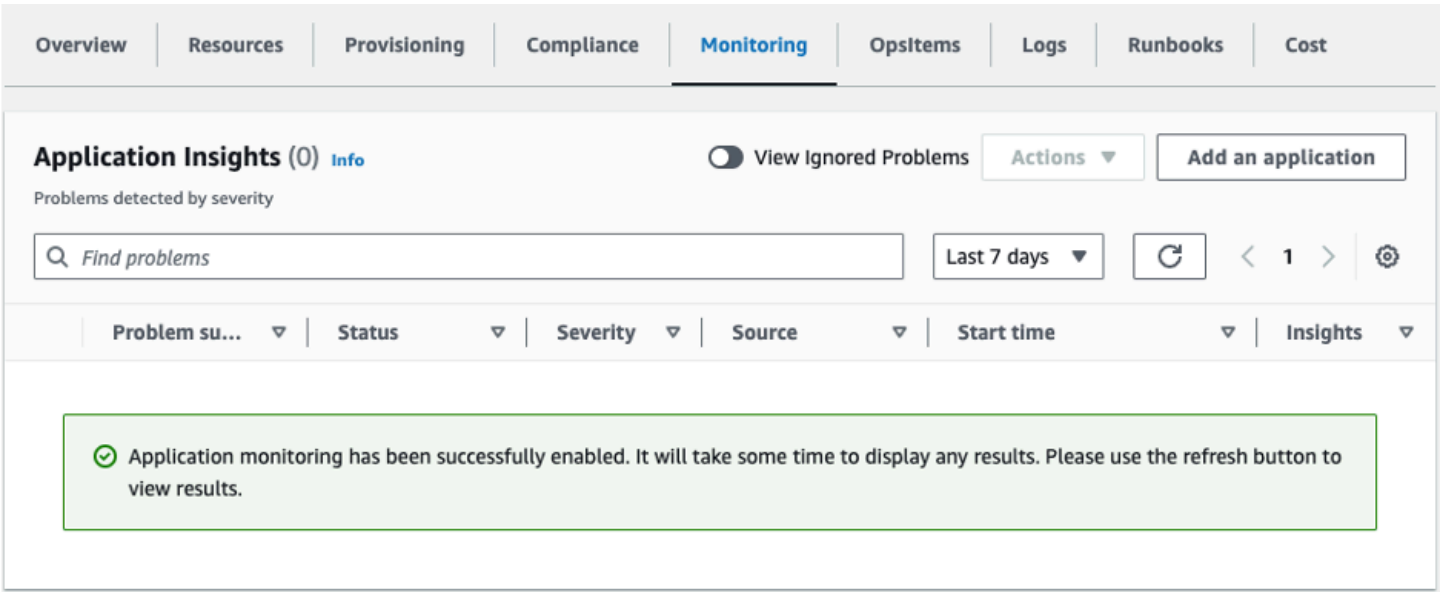
   Le nom de l'application indiquera App Registry dans la colonne Source de l'application et comportera une combinaison du nom de la solution, de la région, de l'ID de compte ou du nom de la pile.
4. Dans l'arborescence des composants, choisissez la pile d'applications que vous souhaitez activer.
5. Dans l'onglet Surveillance, dans Application Insights, sélectionnez Configurer automatiquement Application Insights.

Tableau de bord Application Insights indiquant qu'aucun problème n'a été détecté et qu'il est possible de configurer automatiquement.



La surveillance de vos applications est désormais activée et la boîte de statut suivante apparaît :

Tableau de bord Application Insights affichant un message d'activation de la surveillance réussi.



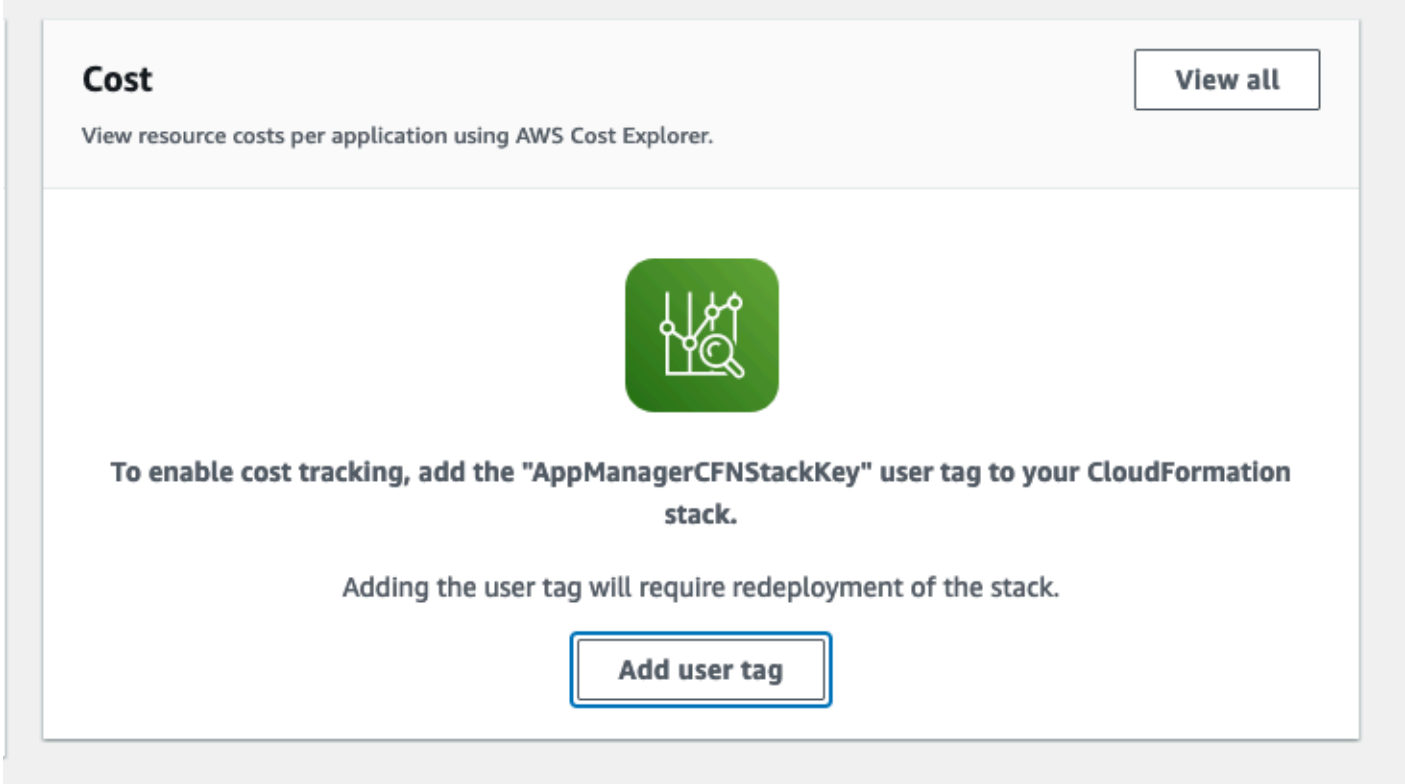
Confirmez les étiquettes de coût associées à la solution

Après avoir activé les balises de répartition des coûts associées à la solution, vous devez confirmer les balises de répartition des coûts pour connaître les coûts de cette solution. Pour confirmer les balises de répartition des coûts :

1. Connectez-vous à la console Systems Manager.
2. Dans le volet de navigation, choisissez Application Manager.
3. Dans Applications, choisissez le nom de l'application pour cette solution, puis sélectionnez-la.

   Le nom de l'application indiquera App Registry dans la colonne Source de l'application et comportera une combinaison du nom de la solution, de la région, de l'ID de compte ou du nom de la pile.
4. Dans l'onglet Aperçu, dans Coût, sélectionnez Ajouter un tag utilisateur.

   Capture d'écran illustrant l'écran d'ajout d'une étiquette utilisateur au coût de l'application



5. Sur la page Ajouter un tag utilisateur, entrez`confirm`, puis sélectionnez Ajouter un tag utilisateur.

Le processus d'activation peut prendre jusqu'à 24 heures et les données du tag peuvent apparaître.

Activer les balises de répartition des coûts associées à la solution

Après avoir activé Cost Explorer, vous devez activer les balises de répartition des coûts associées à cette solution pour connaître les coûts de cette solution. Les balises de répartition des coûts ne peuvent être activées qu'à partir du compte de gestion de l'organisation. Pour activer les balises de répartition des coûts :

1. Connectez-vous à la console AWS Billing and Cost Management and Cost Management.
2. Dans le volet de navigation, sélectionnez Balises de répartition des coûts.
3. Sur la page Balises de répartition des coûts, filtrez le tag AppManager CFNStack Key, puis sélectionnez le tag parmi les résultats affichés.
4. Choisissez Activer.

AWS Cost Explorer

Vous pouvez consulter l'aperçu des coûts associés à l'application et aux composants de l'application dans la console Application Manager grâce à l'intégration à AWS Cost Explorer, qui doit d'abord être activé. Cost Explorer vous aide à gérer les coûts en fournissant une vue des coûts et de l'utilisation de vos ressources AWS au fil du temps. Pour activer Cost Explorer pour la solution :

1. Connectez-vous à la console AWS Cost Management.
2. Dans le volet de navigation, sélectionnez Cost Explorer pour visualiser les coûts et l'utilisation de la solution au fil du temps.

Mettre à jour la solution

Si vous avez déjà déployé la solution, suivez cette procédure pour mettre à jour le CloudFormation stack de la solution afin de bénéficier des dernières fonctionnalités et améliorations. Le processus de mise à niveau comporte trois parties :

- Étape 1 : Mettre à jour le tableau de bord de déploiement
- Étape 2 : migrer les configurations de cas d'utilisation
- Étape 3 : Mettre à jour les cas d'utilisation

> **Note**
>
> 1. Dans la version 2.0.0, l'intégration avec Anthropic et Hugging Face est devenue obsolète au profit d'Amazon Bedrock et Amazon AI. SageMaker Vous pouvez déployer des modèles disponibles via Hugging Face SageMaker JumpStart. Reportez-vous à la section Utiliser Hugging Face avec SageMaker Amazon AI pour plus de détails.
> 2. Assurez-vous de tester le processus de mise à jour dans un environnement hors production avant d'exécuter ces étapes.

Étape 1 : Mettre à jour le tableau de bord de déploiement

1. Connectez-vous à la CloudFormation console, sélectionnez votre CloudFormation stack existant, puis sélectionnez Mettre à jour.
2. Sélectionnez Remplacer le modèle actuel.
3. Sous Spécifier le modèle :
   a. Sélectionnez l'URL Amazon S3.
   b. Copiez le dernier lien du CloudFormation modèle.
   c. Collez le lien dans le champ URL d'Amazon S3.
   d. Vérifiez que l'URL du modèle s'affiche correctement dans la zone de texte URL Amazon S3, puis choisissez Next. Choisissez Suivant à nouveau.
4. Sous Paramètres, passez en revue les paramètres du modèle et modifiez-les si nécessaire. Pour plus de détails sur les paramètres, voir Étape 1 : Lancer la pile du tableau de bord de déploiement.

5. Choisissez Next (Suivant).
6. Sur la page Configurer les options de pile, choisissez Suivant.
7. Sur la page Vérification, vérifiez et confirmez les paramètres. Cochez la case indiquant que le modèle créera des ressources IAM.
8. Choisissez Afficher l'ensemble de modifications et vérifiez les modifications.
9. Choisissez Mettre à jour la pile pour déployer la pile.

Vous pouvez consulter l'état de la pile dans la CloudFormation console AWS dans la colonne Status. Vous devriez recevoir le statut UPDATE_COMPLETE dans environ 10 minutes.

Si la version existante de la solution était antérieure à la version 2.0.0, la mise à jour créera une pile d'interface utilisateur Web (qui remplace l'`amplify-ui`implémentation de l'écran de connexion par une interface utilisateur hébergée par Cognito) et une nouvelle CloudFront URL, qui peut être obtenue dans la section Output de la CloudFormation console une fois que le statut de la pile est UPDATE_COMPLETE.

> **Note**
>
> Les cas d'utilisation existants créés à l'aide de versions antérieures à la version 2.0.0 ne seront PAS affichés tant que vous n'aurez pas effectué les étapes décrites ci-dessous.

Étape 2 : Migrer les configurations de cas d'utilisation (uniquement les mises à jour provenant de versions inférieures à 2.0.0)

Le schéma de stockage et la configuration des cas d'utilisation du service AWS pour le stockage ont changé dans la version 2.0.0. Suivez les étapes décrites dans le guide de l'utilisateur de la migration GAAB v2 à l'aide du script gaab_v2_migration.py. Après avoir exécuté le script, vous pouvez accéder au tableau de bord de déploiement pour consulter les cas d'utilisation déployés.

> **Note**
>
> Vous devez suivre les étapes ci-dessous pour terminer la migration des cas d'utilisation.

Étape 3 : Mettre à jour les cas d'utilisation

Vous pouvez modifier les cas d'utilisation déployés avec les nouvelles fonctionnalités disponibles dans les dernières versions de GAAB. Voir Utiliser la solution pour plus d'informations sur l'utilisation des fonctionnalités de cette solution.

Pour mettre à jour les cas d'utilisation vers la dernière version, vous devez suivre les étapes « Modifier » les cas d'utilisation dans le tableau de bord de déploiement (bien que vous ne puissiez apporter aucune modification). Cette action déclenche une mise à jour de la CloudFormation pile avec la dernière version du modèle.

> **Note**
>
> Les cas d'utilisation créés avec les versions 1.x ou 2.x de la solution peuvent ne pas fonctionner avec les versions ultérieures. Par conséquent, nous recommandons de cloner les cas d'utilisation existants créés avec des versions antérieures à la version 3.0.0 via le tableau de bord de déploiement. Ensuite, migrez progressivement et remplacez-le par de nouveaux cas d'utilisation créés à l'aide de la version 3.0.0 ou ultérieure.

Résolution des problèmes

Cette section fournit des instructions de dépannage pour le déploiement et l'utilisation de la solution.

Si ces instructions ne répondent pas à votre problème, Contacter le Support fournit des instructions pour ouvrir un dossier d'assistance pour cette solution.

Problème : le déploiement d'une configuration compatible VPC, avec Create a VPC for me, échoue

La pile du tableau de bord de déploiement ou la pile de cas d'utilisation échoue car elle n' CloudFormation a pas pu provisionner les ressources réseau VPC.

Résolution

Vérifiez les limites de quota pour VPCs et Elastic IPs dans votre compte. Les limites par défaut sont de 5 pour Elastic IPs et VPCs par compte AWS, par région AWS.

> **Note**
>
> Lorsque la solution crée un VPC, un seul déploiement compatible VPC (tableau de bord de déploiement ou cas d'utilisation) est un déploiement 2-AZ avec 1 sous-réseau public et 1 sous-réseau privé dans chaque zone de zone, chaque sous-réseau public déployant une passerelle NAT. Avec 2 passerelles NAT, le déploiement consomme 2 adresses IP publiques par rapport à la limite de quota.

Quelques limites à connaître (par compte, par région) :

- Nombre de VPCs - 5
- Nombre d'adresses IP publiques : 5
- Nombre de points de terminaison VPC Gateway : 20
- Nombre de points de terminaison VPC d'interface : 20

Problème : la pile de cas d'utilisation ne peut pas être supprimée une CloudFormation fois la pile du tableau de bord de déploiement supprimée

Si la pile du tableau de bord de déploiement est supprimée CloudFormation avant que toutes les piles de cas d'utilisation ne soient supprimées, les cas d'utilisation peuvent se retrouver verrouillés (inutilisables). Cela est dû au fait qu'un rôle IAM créé par la pile du tableau de bord de déploiement n'existe plus, ce qui empêche toute modification de la pile de cas d'utilisation.

Résolution

> **Warning**
>
> Assurez-vous de nettoyer tous les rôles créés manuellement immédiatement après leur utilisation. Il s'agit d'autorisations élevées que les utilisateurs peuvent exploiter pour élever des rôles.

Recréez le rôle IAM supprimé pour permettre la suppression des CloudFormation piles :

1. Ouvrez la CloudFormation console et déterminez le rôle associé à votre pile verrouillée.
   a. L'ARN du rôle se trouve dans la section d'informations sur la pile intitulée Rôle IAM.
   b. Le nom du rôle est celui qui suit après:role/ dans l'ARN du rôle IAM (par exemple, arn:aws:iam : ::role/) <account-id><role-name>
2. Créez un nouveau rôle dans IAM portant le même nom que le rôle supprimé.
   a. Sélectionnez le service AWS comme entité de confiance et sélectionnez-le dans le CloudFormationmenu déroulant.
   b. Ajoutez les autorisations nécessaires. Si vous n'êtes pas sûr des autorisations requises, vous pouvez utiliser la AdministratorAccesspolitique gérée par AWS.
   c. Entrez le nom du rôle exactement tel qu'il a été obtenu à l'étape 1.
3. Retournez à la CloudFormation console et supprimez les piles verrouillées.
4. Une fois que toutes les piles verrouillées ont été supprimées avec succès, retournez dans IAM et supprimez tous les rôles créés à l'étape 2.

Problème : l'interface utilisateur du cas d'utilisation ne reflète pas les modifications apportées aux paramètres

Lorsque les cas d'utilisation sont mis à jour, l'interface utilisateur est déployée sur CloudFront. Cependant, étant donné que CloudFront les déploiements sont mis en cache ainsi que le fichier de configuration qui dicte la manière dont certains paramètres sont présentés à l'utilisateur, ces modifications peuvent ne pas être prises en compte immédiatement.

Résolution

La CloudFront distribution peut être invalidée pour forcer la propagation de la nouvelle configuration aux utilisateurs du frontend.

1. Ouvrez la CloudFormation console et déterminez la CloudFront distribution associée à votre pile de cas d'utilisation.
   a. La pile de cas d'utilisation doit commencer par le même nom que celui que vous avez utilisé lors du déploiement du cas d'utilisation.
   b. Localisez la pile imbriquée correspondant à l'interface utilisateur. Le nom de la pile imbriquée doit commencer par WebAppS3 UINested UINested StackResource StackS3.
   c. Dans l'onglet Ressources, recherchez le type de ressource AWS::CloudFront::Distribution, puis sélectionnez l'identifiant physique. Cela ouvrira la distribution dans la CloudFront console.
2. Accédez à l'onglet Invalidations, puis choisissez Créer une invalidation et entrez le chemin /*. Cela invalidera tous les chemins.
3. Dans votre propre navigateur, supprimez tous les cookies et fichiers mis en cache liés au cas d'utilisation.

Contacter AWS Support

Si vous disposez d'AWS Business Support+, d'AWS Enterprise Support ou d'Unified Operations, vous pouvez utiliser le centre de support AWS pour obtenir l'assistance d'experts concernant cette solution. Les sections suivantes fournissent des instructions.

Créer un dossier

1. Connectez-vous au Centre de Support.
2. Choisissez Create case (Créer une demande).

Comment pouvons-nous vous aider ?

1. Choisissez Technique.
2. Dans le champ Service, sélectionnez Solutions.
3. Dans Catégorie, sélectionnez Autres solutions.
4. Pour Severity, sélectionnez l'option qui correspond le mieux à votre cas d'utilisation.
5. Lorsque vous entrez le service, la catégorie et la gravité, l'interface contient des liens vers des questions de dépannage courantes. Si vous ne parvenez pas à résoudre votre question à l'aide de ces liens, sélectionnez Étape suivante : Informations supplémentaires.

Informations supplémentaires

1. Dans le champ Objet, saisissez un texte résumant votre question ou problème.
2. Pour la description, décrivez le problème en détail, notamment le nom de cette solution : Generative AI Application Builder on AWS.
3. Choisissez Joindre des fichiers.
4. Joignez les informations dont AWS Support a besoin pour traiter la demande.

Aidez-nous à résoudre votre cas plus rapidement

1. Entrez les informations demandées.
2. Cliquez sur Étape suivante : résoudre maintenant ou nous contacter.

Résolvez maintenant ou contactez-nous

1. Passez en revue les solutions Solve now.
2. Si vous ne parvenez pas à résoudre votre problème avec ces solutions, choisissez Contactez-nous, entrez les informations demandées, puis choisissez Soumettre.

Désinstallez la solution

> **Note**
>
> Les déploiements créés via le tableau de bord de déploiement ne sont pas destinés à être gérés en dehors de la solution. Assurez-vous de supprimer et de nettoyer tous les déploiements depuis le tableau de bord de déploiement, avant de supprimer le stack. CloudFormation

Vous pouvez désinstaller la solution Generative AI Application Builder on AWS depuis la console de gestion AWS ou en utilisant l'interface de ligne de commande AWS. Vous devez supprimer manuellement les compartiments Amazon S3, les index Amazon Kendra CloudWatch ou les journaux créés par cette solution. Les solutions AWS ne suppriment pas automatiquement les compartiments Amazon S3, les index Amazon Kendra CloudWatch ou les journaux si vous avez stocké des données à conserver.

Utilisation de la AWS Management Console

1. Connectez-vous à la CloudFormation console AWS.
2. Sur la page Stacks, sélectionnez la pile d'installation de cette solution.
3. Sélectionnez Delete (Supprimer).

Utilisation de l'interface de ligne de commande AWS

Déterminez si l'interface de ligne de commande AWS (AWS CLI) est disponible dans votre environnement. Pour les instructions d'installation, consultez la section Qu'est-ce que l'interface de ligne de commande AWS dans le guide de l'utilisateur de l'interface de ligne de commande AWS. Après avoir confirmé que l'AWS CLI est disponible, exécutez la commande suivante.

```
$ aws cloudformation delete-stack --stack-name <installation-stack-name>
```

Étapes de désinstallation manuelle

Suppression des compartiments Amazon S3

Cette solution est configurée pour conserver le compartiment Amazon S3 créé par la solution si vous décidez de supprimer la CloudFormation pile AWS afin d'éviter toute perte de données accidentelle. Après avoir désinstallé la solution, vous pouvez supprimer manuellement ce compartiment Amazon S3 si vous n'avez pas besoin de conserver les données. Suivez ces étapes pour supprimer le compartiment Amazon S3.

1. Connectez-vous à la console Amazon S3.
2. Dans le volet de navigation, sélectionnez Buckets.
3. Localisez les <stack-name>compartiments S3.
4. Sélectionnez le compartiment S3, puis choisissez Supprimer.

Pour supprimer le compartiment S3 à l'aide de l'interface de ligne de commande AWS, exécutez la commande suivante. Il n'est pas nécessaire de vider le compartiment au préalable lorsque vous utilisez l'option --force.

```
$ aws s3 rb s3://<bucket-name> --force
```

Supprimer les index Amazon Kendra

Pour éviter toute perte de données accidentelle, cette solution est configurée pour conserver les index Amazon Kendra créés par la solution lorsque la pile CloudFormation AWS a été supprimée. Après avoir désinstallé la solution, vous pouvez supprimer manuellement les index Amazon Kendra dont vous n'avez plus besoin de conserver les données. Suivez ces étapes pour supprimer l'index Amazon Kendra.

1. Connectez-vous à la console Amazon Kendra.
2. Dans le volet de navigation, sélectionnez Indexes.
3. Localisez et sélectionnez l'index que vous souhaitez supprimer.
4. Choisissez Supprimer pour supprimer l’index sélectionné.

Pour supprimer l'index Amazon Kendra à l'aide de l'AWS CLI, exécutez la commande suivante :

```
$ aws kendra delete-index --id<index-id>
```

Supprimer les CloudWatch journaux

Pour éviter toute perte de données accidentelle, nous avons configuré cette solution pour conserver les CloudWatch journaux si vous décidez de supprimer la CloudFormation pile. Après avoir désinstallé la solution, vous pouvez supprimer manuellement les journaux si vous n'avez pas besoin de conserver les données. Procédez comme suit pour supprimer les CloudWatch journaux.

1. Connectez-vous à la CloudWatch console Amazon.
2. Dans le volet de navigation, sélectionnez Log Groups.
3. Localisez les groupes de journaux créés par la solution.
4. Sélectionnez l'un des groupes de journaux.
5. Choisissez Actions, puis Delete (Supprimer).

Répétez les étapes jusqu'à ce que vous ayez supprimé tous les groupes de journaux de solutions.

Utilisez la solution

Accès à l'interface utilisateur

Pendant le processus de déploiement de la pile (pour le tableau de bord de déploiement et les cas d'utilisation), un e-mail est envoyé à l'adresse e-mail configurée. L'e-mail contient les informations d'identification temporaires de l'utilisateur qu'il peut utiliser pour s'inscrire et accéder à l'interface Web.

> **Note**
>
> L' DevOps utilisateur ayant accès à l'AWS Management Console doit fournir à l'administrateur l' CloudFront URL de l'interface utilisateur du tableau de bord de déploiement lorsque la pile est terminée.

Pour les cas d'utilisation, l'utilisateur administrateur ayant accès à l'interface utilisateur du tableau de bord de déploiement doit fournir à l'utilisateur professionnel l' CloudFront URL de l'interface utilisateur du cas d'utilisation une fois le déploiement terminé.

Une fois connecté, l'utilisateur peut interagir avec la solution UIs, soit le tableau de bord de déploiement dans le cas des administrateurs, soit le cas d'utilisation dans le cas des utilisateurs professionnels.

Comment mettre à jour un déploiement

Sur la page d'accueil du tableau de bord de déploiement (ou sur la page de détails d'un déploiement), vous pouvez modifier la configuration utilisée par un déploiement. Vous ne pouvez modifier que les déploiements dont le statut est CREATE_COMPLETE ou UPDATE_COMPLETE.

À l'exception du nom du cas d'utilisation, toutes les autres options sont modifiables pour un déploiement. Modifiez simplement les valeurs que vous souhaitez modifier et redéployer.

En fonction de l'étendue des modifications apportées, le délai de redéploiement varie. Cela peut prendre quelques secondes si des paramètres simples ont changé (par exemple, les paramètres du modèle), ou plus de 30 minutes si des options liées à une infrastructure plus importante ont changé (par exemple, demande de création de l'index Amazon Kendra pour le cas d'utilisation du texte RAG).

Une fois la modification terminée avec succès, le statut de l'application indiquera un statut UPDATE_COMPLETE. À ce stade, vous pouvez accéder à l'interface utilisateur déployée via l' CloudFront URL et interagir avec le déploiement modifié.

> **Note**
>
> Il peut être plus facile d'exécuter plusieurs déploiements side-by-side si vous souhaitez comparer différents paramètres ou LLMs. Utilisez la fonction Clone pour utiliser rapidement une configuration existante afin de lancer un nouveau déploiement.

Comment cloner un déploiement

Sur la page d'accueil du tableau de bord des déploiements (ou sur la page de détails d'un déploiement), vous pouvez cloner la configuration utilisée par un déploiement. Le clonage d'un déploiement lance l'assistant de déploiement d'un nouveau cas d'utilisation, mais la plupart des champs sont préremplis avec les mêmes valeurs.

Il s'agit d'une opération pratique qui vous permet de dupliquer rapidement des déploiements dont les paramètres ont été modifiés, de relancer un déploiement supprimé ou d'en comparer plusieurs LLMs dans le cadre de déploiements par ailleurs identiques.

Comment supprimer un déploiement

Sur la page d'accueil du tableau de bord des déploiements (ou sur la page de détails d'un déploiement), vous pouvez le supprimer une fois que vous n'en avez plus besoin. La suppression d'un déploiement entraîne une opération de suppression de CloudFormation pile et déprovisionne les ressources nécessaires au déploiement.

Par défaut, un déploiement supprimé reste sur le tableau de bord pour activer la fonctionnalité de clonage. Pour supprimer complètement un déploiement du tableau de bord afin qu'il cesse d'être suivi dans l'interface utilisateur, choisissez Supprimer définitivement dans la fenêtre de confirmation de suppression.

> **Important**
>
> Certaines ressources sont laissées pour compte lors de la suppression de la pile et doivent être supprimées manuellement. Reportez-vous à la section Désinstallation manuelle pour plus de détails sur les ressources conservées et sur la manière de les nettoyer.

Configuration d'un modèle linguistique étendu (LLM)

Le choix du LLM adapté à votre cas d'utilisation dépend d'un large éventail de facteurs spécifiques à vos besoins et au type d'expérience client que vous souhaitez créer. Cette solution ne semble pas prescriptive, mais vise plutôt à vous fournir les outils nécessaires pour évaluer ce qui fonctionne le mieux pour votre application.

L'espace généré par l'IA évolue rapidement. Il vous incombe donc de vous tenir au courant des derniers modèles, des techniques d'optimisation et des meilleures pratiques afin de vous assurer de créer les bonnes expériences pour vos clients.

> **Note**
>
> Si vous travaillez avec des données non publiques ou sensibles, veillez à sélectionner une option LLM à l'aide des services AWS (tels qu'Amazon Bedrock ou Amazon SageMaker AI). Cela améliore le niveau de sécurité global de votre déploiement en conservant les données dans votre région et sur le réseau AWS par rapport à l'utilisation d'un LLM hébergé par un fournisseur tiers.

Utiliser Amazon SageMaker AI en tant que fournisseur de LLM

Depuis la version v1.3.0, Amazon SageMaker AI est disponible en tant que fournisseur modèle pour les cas d'utilisation de Text. Cette fonctionnalité vous permet d'utiliser un point de terminaison d'inférence SageMaker AI déjà existant dans le compte AWS de la solution. Voici quelques conseils pour commencer.

> **Important**
>
> La solution ne gère pas le cycle de vie de vos terminaux d' SageMaker IA. Vous êtes responsable de la suppression des points de terminaison SageMaker AI une fois qu'ils ne sont plus nécessaires pour ne plus entraîner de frais supplémentaires.

Création d'un point de terminaison SageMaker AI

Vous pouvez utiliser Amazon SageMaker AI JumpStart pour déployer rapidement un point de terminaison.

Vous pouvez également utiliser un point de terminaison d' SageMaker IA basé sur la génération de texte et le déployer à l'aide du service d' SageMaker IA de base. Reportez-vous à la JumpStart documentation de l'SageMaker IA pour obtenir un guide étape par étape sur le déploiement d'un modèle à des fins d'inférence.

> **Note**
>
> models/LLMs Les fondations sont généralement assez volumineuses et peuvent souvent nécessiter l'utilisation de grandes instances de calcul accéléré. La plupart de ces instances de plus grande taille ne sont peut-être pas disponibles par défaut dans votre compte AWS. Reportez-vous aux quotas d'SageMaker IA par défaut et assurez-vous de demander une augmentation de quota avant le déploiement afin d'éviter les échecs de déploiement courants.

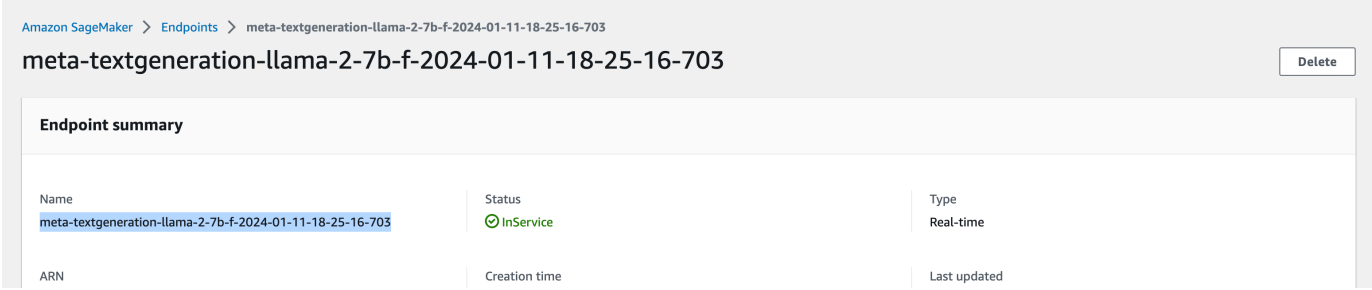
Utiliser un point de terminaison SageMaker AI pour créer un déploiement de cas d'utilisation textuel

Pour déployer un nouveau cas d'utilisation du texte à l'aide d'un point de terminaison SageMaker AI à des fins d'inférence :

1. Créez un nouveau cas d'utilisation via l'assistant du tableau de bord de déploiement et complétez les formulaires jusqu'à ce que vous atteigniez la page de sélection des modèles.
2. Sur la page Modèles, sélectionnez SageMaker AI comme fournisseur de modèles. Cela générera un formulaire personnalisé nécessitant la saisie de trois éléments clés par l'utilisateur :

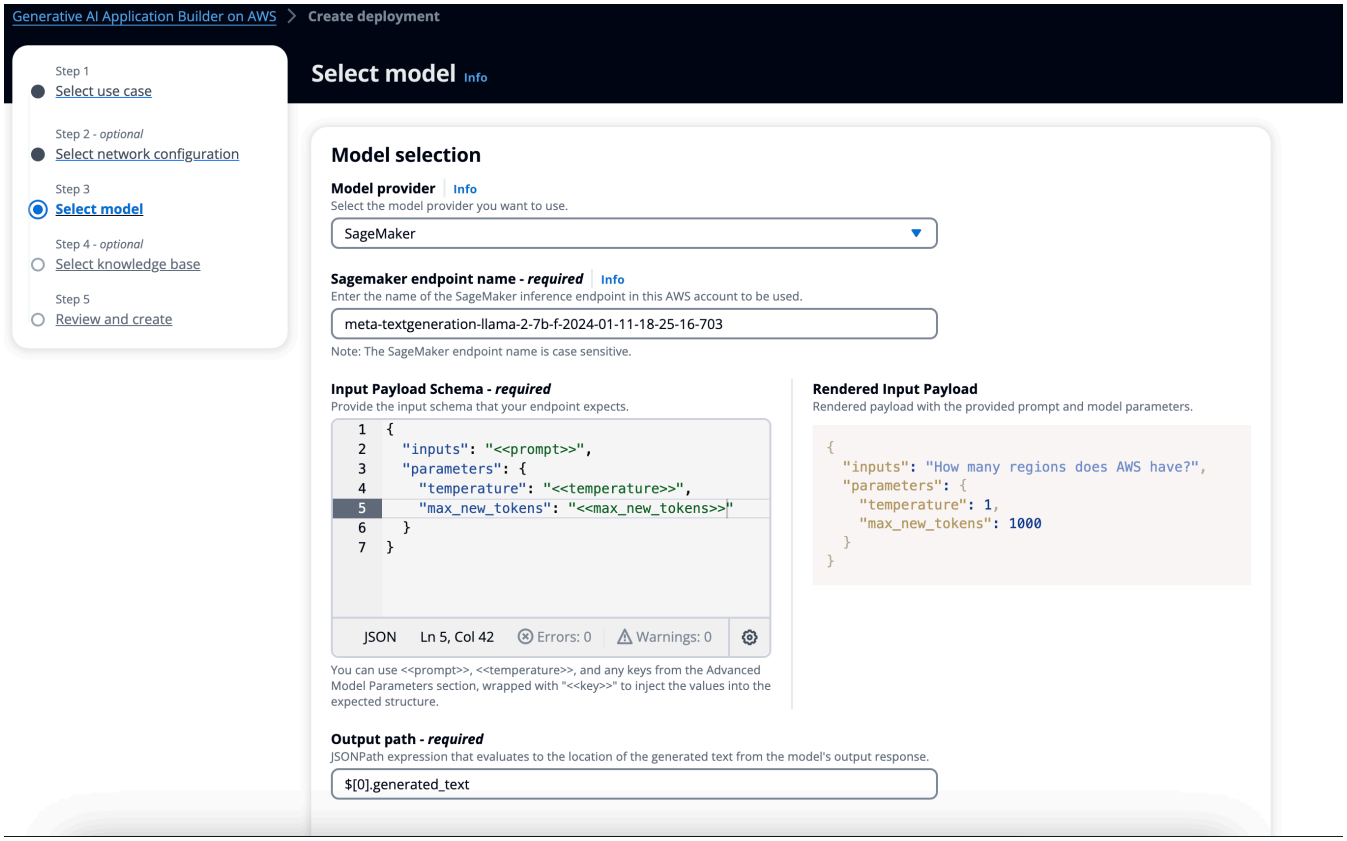
- Le nom du point de terminaison SageMaker AI que vous souhaitez utiliser. DevOps les utilisateurs peuvent l'obtenir depuis la console AWS. Notez que le point de terminaison doit se trouver dans le même compte et dans la même région que ceux dans lesquels la solution est déployée.

  Emplacement du nom du point de terminaison sur la console AWS



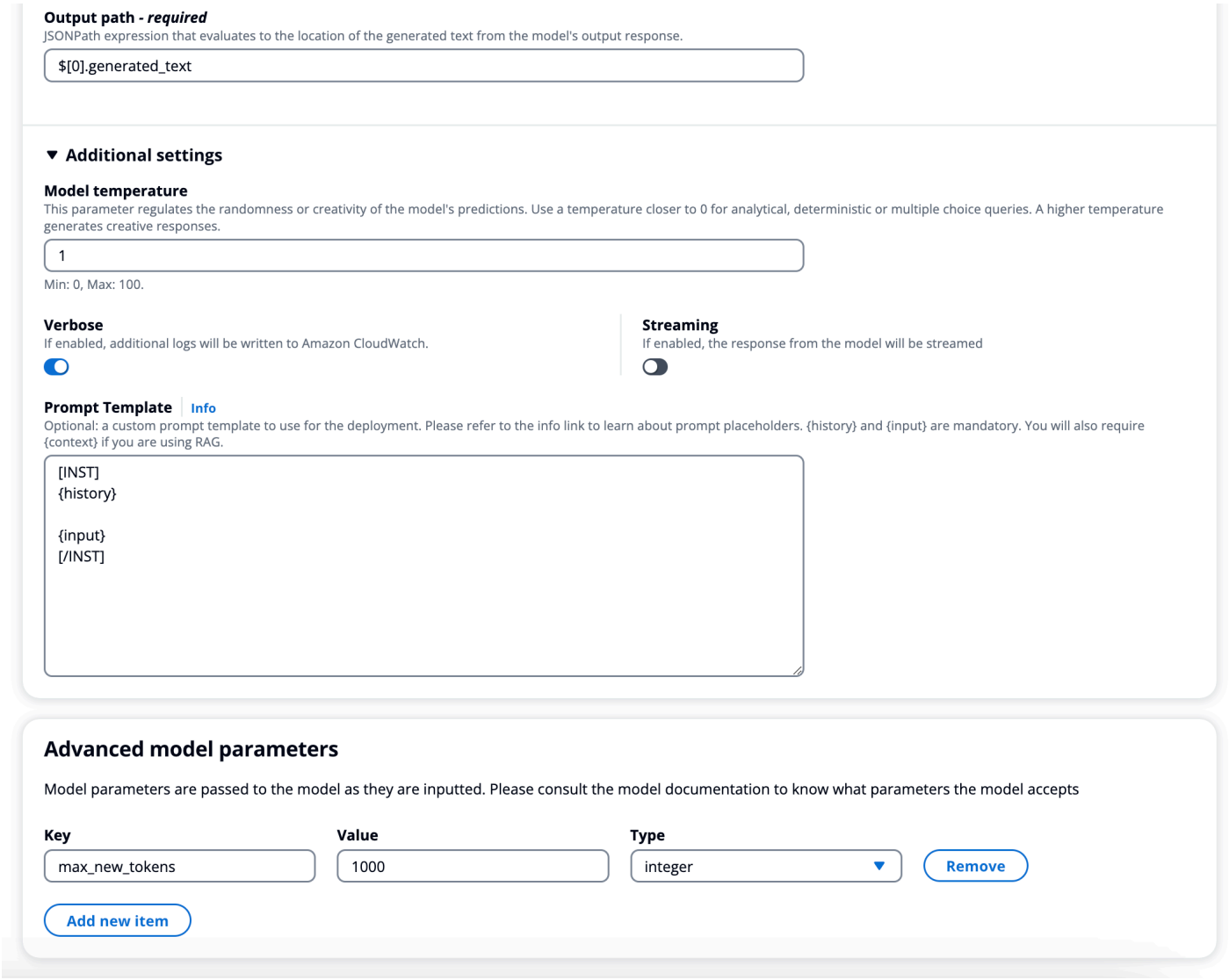
- Schéma de la charge utile d'entrée attendue par le point de terminaison. Pour prendre en charge le plus grand nombre de points de terminaison, les utilisateurs administrateurs doivent indiquer à la solution comment leur point de terminaison prévoit que l'entrée sera formatée. Dans l'assistant de sélection du modèle, fournissez le schéma JSON pour la solution à envoyer au point de terminaison. Vous pouvez ajouter des espaces réservés pour injecter des valeurs statiques et dynamiques dans la charge utile de la demande. Les options disponibles sont les suivantes :
  - Espaces réservés obligatoires : \ < \ <prompt \ > \ > seront remplacés dynamiquement par l'entrée complète (par exemple, historique, contexte et saisie utilisateur conformément au modèle d'invite) à envoyer au point de terminaison SageMaker AI lors de l'exécution.
  - Des espaces réservés facultatifs : \ < \ <temperature \ > \ > *, \ * ainsi que tous les paramètres définis dans les paramètres avancés du modèle peuvent être fournis au point de terminaison. Toute chaîne contenant un espace réservé entre \ < \ < and \ > \ > (par exemple, \ < \ <max_new_tokens \ > \ >) sera remplacée par la valeur du paramètre de modèle avancé du même nom.

    Exemple de schéma de saisie : définition des champs obligatoires, de l'invite et de la température, ainsi qu'un paramètre avancé personnalisé, max_new_tokens. Le chemin de sortie doit être fourni sous forme de JSONPath chaîne valide



3. Emplacement de la réponse sous forme de chaîne LLMs générée dans la charge utile de sortie. Cela doit être fourni sous forme d' JSONPath expression pour indiquer où la réponse textuelle finale présentée aux utilisateurs devrait être accessible depuis l'objet de retour et la réponse du point de terminaison.

   Exemple d'ajout de paramètres de modèle avancés à utiliser dans le schéma d'entrée SageMaker AI (voir Figure 2 pour les options/paramètres précédents)



> **Note**
>
> SageMaker L'IA prend désormais en charge l'hébergement de plusieurs modèles derrière le même point de terminaison, et il s'agit de la configuration par défaut lors du déploiement d'un point de terminaison dans la version actuelle d' SageMaker AI Studio (et non dans Studio Classic).
>
> Si votre point de terminaison est configuré de cette manière, vous devrez ajouter InferenceComponentNameà la section des paramètres avancés du modèle une valeur correspondant au nom du modèle que vous souhaitez utiliser.

Paramètres LLM avancés

Lorsque vous utilisez Amazon Bedrock, vous pouvez configurer certains paramètres avancés pour vos modèles, tels qu'Amazon Bedrock Guardrails, Provisioned Throughput pour Amazon Bedrock et des paramètres de modèle supplémentaires.

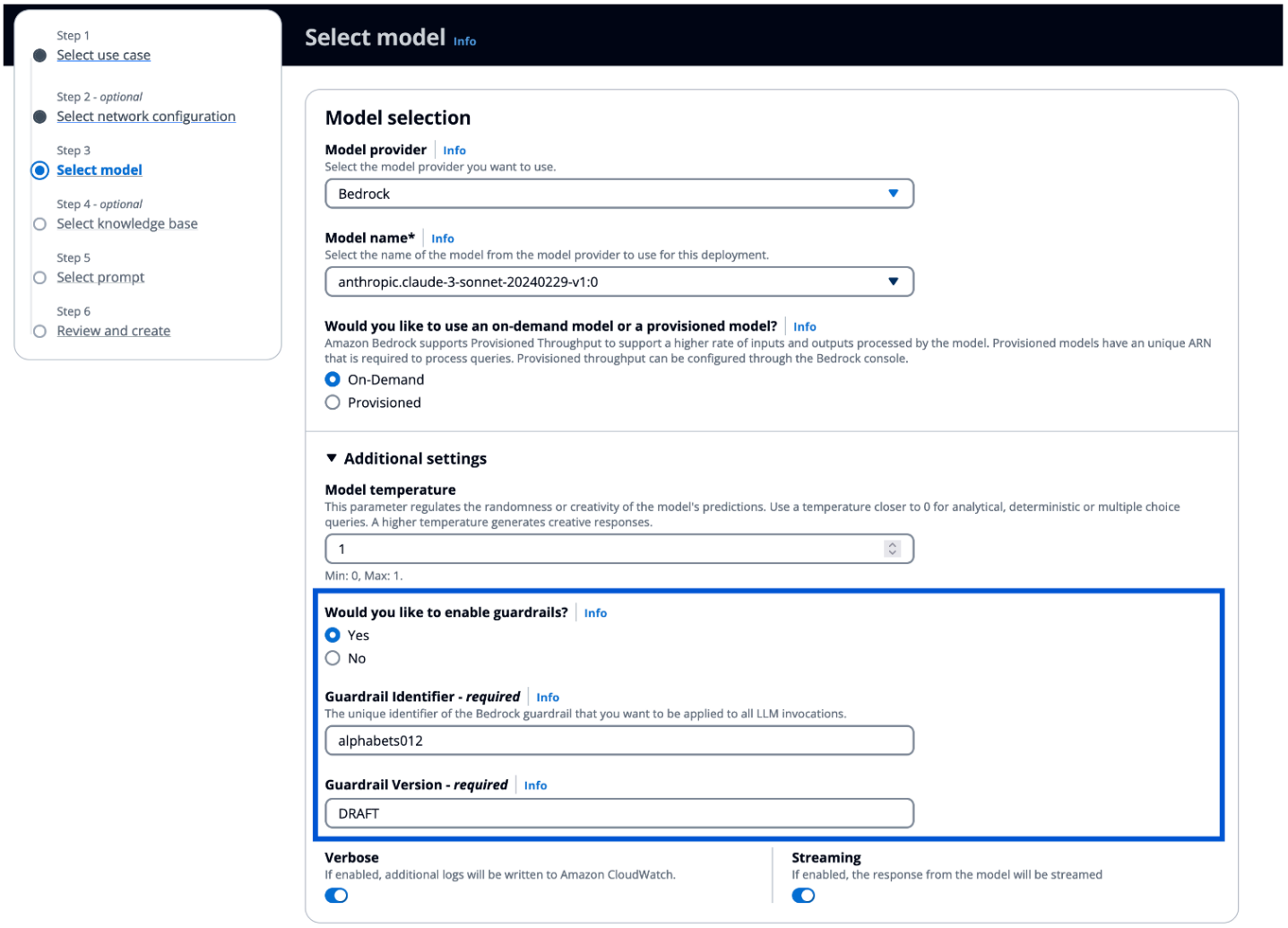
Barrières de protections Amazon Bedrock

Amazon Bedrock Guardrails est une fonctionnalité d'Amazon Bedrock qui évalue les entrées des utilisateurs et les réponses LLM en fonction des politiques configurées par l'utilisateur et fournit un niveau de protection supplémentaire, quel que soit le LLM sous-jacent sélectionné par l'utilisateur pour un cas d'utilisation. Un garde-corps comprend deux politiques visant à éviter les contenus appartenant à des catégories indésirables ou nuisibles :

1. Sujets refusés pour définir un ensemble de sujets indésirables dans le contexte de la demande de l'utilisateur, par exemple, les conseils d'investissement dans une application financière, et,
2. Filtres de contenu****qui permettent de filtrer les demandes des utilisateurs ou les modèles de réponses contenant du contenu préjudiciable.

Pour une utilisation dans la solution Generative AI Application Builder, un garde-corps doit être configuré dans la console Amazon Bedrock à l'aide de l'assistant de création de garde-corps. Une fois créé, vous pouvez ajouter ce garde-corps à votre cas d'utilisation du chat créé à l'aide de l'assistant de solution Generative AI Application Builder dans les paramètres supplémentaires de l'étape de sélection du modèle en fournissant votre identifiant de garde-corps et votre version de garde-corps.

Décrit l'assistant de déploiement permettant d'activer Amazon Bedrock Guardrails



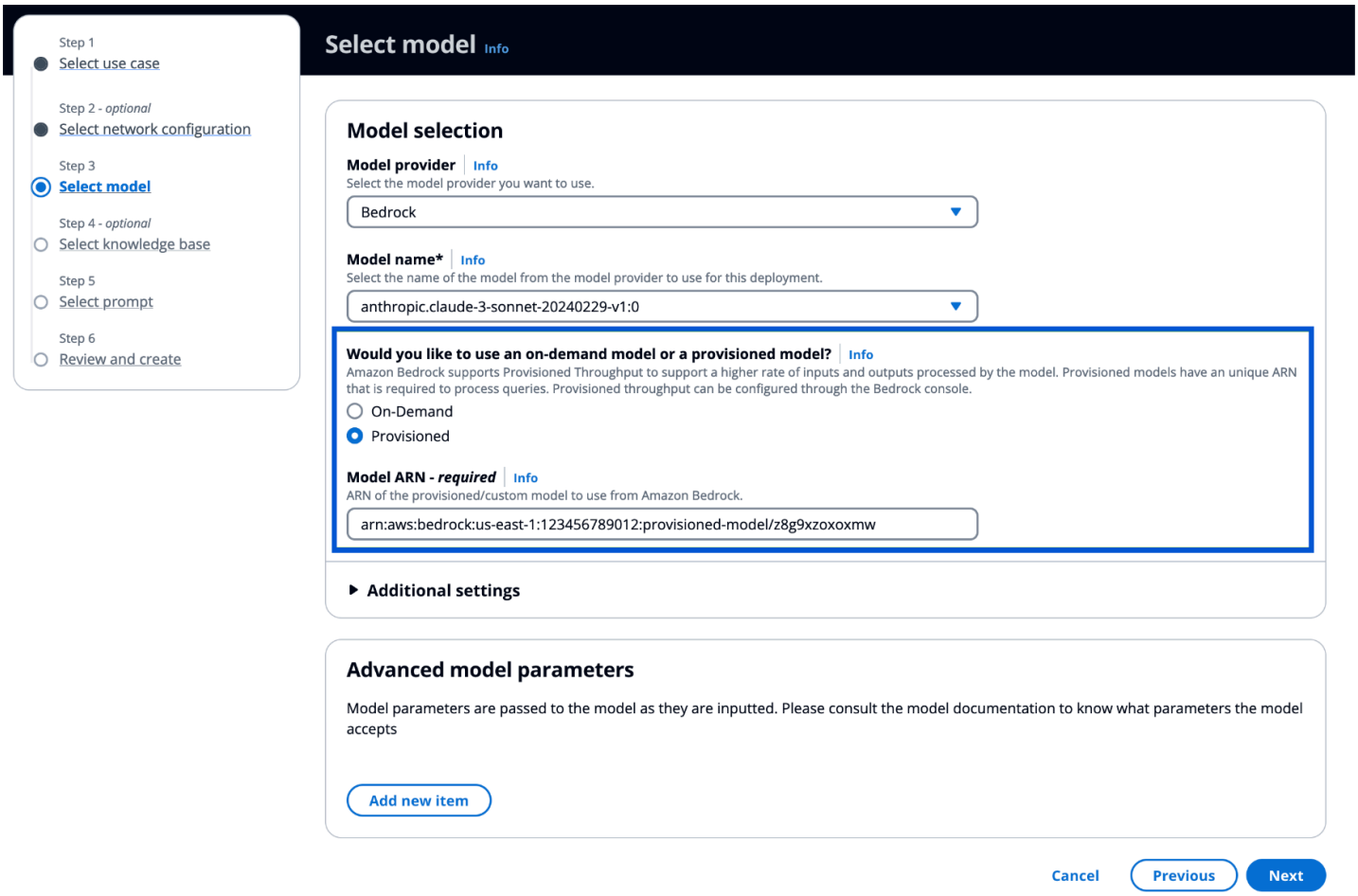
Débit provisionné pour Amazon Bedrock

Chaque modèle Amazon Bedrock à la demande respecte la limite de quota de compte spécifique à la région pour l'inférence du modèle. Par exemple, Anthropic Claude 2.x sur Bedrock autorise actuellement le traitement de 500 demandes et 500 000 jetons par minute dans les régions us-east-1 et us-west-2. Vous pouvez également utiliser la solution avec vos modèles affinés ou préentraînés en continu. Dans de tels cas, Amazon Bedrock autorise un débit provisionné, ce qui permet d'exécuter des charges de travail d'inférence importantes et cohérentes pour votre base, ainsi que des modèles préentraînés affinés ou continus à utiliser dans des applications de production.

Une fois que le Provisioned Throughput est acheté dans la console Amazon Bedrock, un ARN de modèle est généré pour être utilisé. Vous pouvez désormais fournir cet ARN du modèle dans l'assistant Generative AI Application Builder lors de l'étape de sélection du modèle. Pour ce faire, sélectionnez Bedrock comme fournisseur de modèles et le nom du modèle de base utilisé pour

générer cet ARN de modèle provisionné dans la console Amazon Bedrock. Sélectionnez ensuite « Modèle provisionné » lorsque vous choisissez entre les modèles à la demande et provisionnés, et fournissez votre Model ARN.

Décrit l'assistant de déploiement permettant d'activer le débit provisionné pour Amazon Bedrock



> **Note**
>
> Votre garde-fou et votre débit provisionné doivent se trouver dans la même région que le tableau de bord de déploiement déployé et les piles de cas d'utilisation.

Paramètres du modèle

LLMs acceptent souvent un large éventail de paramètres spécifiques à sa mise en œuvre. Les fournisseurs de modèles fournissent souvent de la documentation décrivant l'ensemble des paramètres pris en charge et leurs utilisations.

La solution transmet les paramètres du modèle directement au modèle sous-jacent. Il est donc important de s'assurer que les paramètres sont correctement définis. Reportez-vous à la documentation du fournisseur du modèle pour obtenir les dernières informations sur les paramètres pris en charge.

Configuration d'Agent Builder

Agent Builder propose des options de configuration complètes pour créer des agents d'IA prêts à être utilisés en production. Cette section décrit comment configurer et gérer les déploiements d'Agent Builder.

Configuration rapide du système

L'invite du système définit le comportement, la personnalité et les capacités de votre agent. Pour configurer l'invite du système :

1. Dans l'assistant Agent Builder, accédez à l'étape Configurer l'agent.
2. Modifiez le modèle d'invite du système dans l'éditeur de texte.
3. Incluez des instructions claires pour :
   - Rôle et objectif de l'agent
   - Comment utiliser les outils disponibles (serveurs MCP)
   - Préférences de formatage des réponses
   - Directives comportementales
4. Utilisez le bouton Rétablir les valeurs par défaut pour restaurer le modèle d'origine si nécessaire.

Bonnes pratiques en matière d'instructions aux agents :

- Soyez précis quant aux capacités et aux limites de l'agent
- Fournir des exemples clairs du comportement souhaité
- Incluez des instructions sur l'utilisation des outils et sur le moment de les invoquer
- Définissez les attentes en matière de format de réponse
- Définissez des limites pour le comportement des agents

Intégration au serveur MCP

Les serveurs MCP (Model Context Protocol) permettent aux agents d'accéder aux outils et aux sources de données de l'entreprise. Pour configurer les serveurs MCP, procédez comme suit :

1. À l'étape Configurer l'agent, recherchez la section Serveurs MCP.
2. Sélectionnez l'un des serveurs MCP disponibles dans le menu déroulant.

> **Note**
>
> Les serveurs MCP doivent être configurés et accessibles avant le déploiement de l'agent. L'agent découvrira et utilisera automatiquement les outils exposés par les serveurs MCP configurés. Reportez-vous à la documentation MCP pour la configuration du serveur et des outils.

Réglages de mémoire

Agent Builder fournit deux types de mémoire pour gérer le contexte et les connaissances :

Mémoire à court terme

Activé par défaut pour tous les agents :

- Maintient le contexte des conversations au cours des sessions
- Capture automatiquement les messages des utilisateurs et les réponses des agents
- Organisé par ActorID et SessionId pour une isolation correcte
- Aucune configuration requise

Mémoire à long terme

Fonctionnalité optionnelle permettant de stocker des informations au cours des sessions :

1. À l'étape Configurer l'agent, recherchez la section Configuration de la mémoire.
2. Cliquez sur Activer la mémoire à long terme pour l'activer.
3. Lorsqu'il est activé, l'agent peut :

- Extrayez et stockez les informations importantes au cours des conversations
- Récupérez le contexte pertinent des sessions précédentes
- Acquérir des connaissances sur les préférences et l'historique des utilisateurs

> **Note**
>
> La mémoire à long terme utilise AgentCore la mémoire avec une stratégie de mémoire sémantique et des paramètres de rétention par défaut.

Surveillance des déploiements d'Agent Builder

Agent Builder fournit une surveillance complète via des CloudWatch tableaux de bord et des métriques.

Accès aux CloudWatch tableaux de bord

1. Accédez à la CloudWatch console de votre compte AWS.
2. Sélectionnez Tableaux de bord dans le menu de navigation de gauche.
3. Trouvez le tableau de bord nomméAgentBuilder-<UseCaseId>.
4. Consultez les métriques en temps réel et les données de performance historiques.

Accès aux journaux et analyse

Les journaux de l'agent sont disponibles dans CloudWatch Logs :

1. Accédez à CloudWatch Logs dans la console AWS.
2. Trouvez les groupes de journaux préfixés par/aws/bedrock-agentcore/runtimes/.
3. Utilisez CloudWatch Insights pour interroger et analyser les journaux.
4. Recherchez des demandes IDs ou des modèles d'erreur spécifiques.

Configuration de Workflow Builder

Workflow Builder permet une orchestration multi-agents par le biais d'un agent superviseur qui délègue le travail à des agents Agent Builder spécialisés.

Création d'un flux de travail

1. Accédez au tableau de bord de déploiement
2. Sélectionnez Créer un cas d'utilisation du flux de travail
3. Configurez l'agent de supervision :
   - Nom : nom descriptif du flux de travail
   - Description : Objectif et capacités
   - Système Prompt : instructions pour la délégation et la coordination des agents
   - Modèle : modèle de base pour l'agent superviseur

Bonnes pratiques relatives aux instructions des superviseurs :

- Décrivez clairement quand utiliser chaque agent spécialisé
- Incluez des instructions pour agréger les résultats de plusieurs agents
- Définissez les attentes relatives au formatage des réponses
- Définissez des limites pour le comportement des délégations

Sélection de l'agent

Sélectionnez les agents Agent Builder à inclure en tant qu'agents spécialisés :

1. Cliquez sur Ajouter un agent dans la configuration du flux de travail
2. Parcourez ou recherchez les agents Agent Builder disponibles
3. Consultez les descriptions des agents
4. Sélectionnez les agents à inclure dans le flux de travail

Descriptions des agents

L'agent superviseur utilise les descriptions des agents pour décider à quel agent déléguer. Assurez-vous que les descriptions expliquent clairement :

- Domaine ou capacité spécialisés de l'agent
- Types de tâches gérées par l'agent
- Attentes en matière d'entrées/sorties

Tester les workflows

Après le déploiement :

1. Accédez au flux de travail via le tableau de bord de déploiement
2. Effectuez des tests avec des requêtes qui nécessitent plusieurs agents
3. Surveiller la délégation des agents dans CloudWatch les journaux
4. Réviser la qualité des réponses et les modèles de délégation
5. Ajustez l'invite du superviseur si la délégation n'est pas optimale

Conseils pour gérer les limites des modèles de jetons

Remarque : La solution ne tente pas directement de gérer les limites de jetons imposées par divers LLMs. Testez et assurez-vous que votre message respecte les limites disponibles imposées par le fournisseur de modèles.

Pour vous aider à contrôler la taille des invites, essayez ce qui suit :

1. Familiarisez-vous avec les limites imposées par le modèle que vous souhaitez utiliser. Ces valeurs peuvent varier considérablement d'un modèle à l'autre. Il est donc important de connaître votre budget disponible avant de commencer.
2. Élaborez votre invite initiale en tenant compte de ce budget et réfléchissez au montant que vous souhaitez économiser pour les éléments dynamiques de l'invite. Par exemple, les entrées de l'utilisateur, l'historique des discussions, les extraits de documents, etc.
3. Sur la page de configuration de l'invite, définissez une limite pour la taille de l'historique de suivi afin de limiter le nombre de tours de conversation inclus dans l'invite.
4. Définissez les limites de retour des documents dans l'assistant de configuration de la base de connaissances. Vous devez essayer de trouver le juste équilibre entre fournir au LLM suffisamment de contexte pour effectuer la tâche, mais pas au point de dépasser les limites de jetons ou d'affecter négativement la latence.
5. Laisse un peu de tampon. N'établissez pas de budget pour un cas typique, réfléchissez aux cas extrêmes, tels que les longues requêtes de saisie, les extraits de documents volumineux ou les longues conversations, et expérimentez avec eux.

Étapes pour créer une image Docker du serveur MCP

Pour utiliser les serveurs MCP (Model Context Protocol) avec Generative AI Application Builder sur AWS, vous devez d'abord créer une image Docker et la stocker dans un référentiel Amazon ECR privé.

> **Note**
>
> À l'heure actuelle, les serveurs MCP déployés dans Amazon Bedrock AgentCore Runtime ne peuvent pas être exportés vers GAAB. Pour que les serveurs MCP soient connectés à des agents créés via GAAB, ils doivent être créés via GAAB.

Étape 1 : Créez votre serveur MCP

Tout d'abord, vous devez avoir prêt à implémenter votre serveur MCP. Pour obtenir des instructions détaillées sur la création d'un serveur MCP, reportez-vous au manuel Amazon Bedrock AgentCore Developer Guide - Create an MCP server.

Nous recommandons la structure de projet suivante :

```
.
### __init__.py
### extras/
#   ### extra_dependencies.py
#   ### Dockerfile
### requirements.txt
### server.py <-- Server Entry point
```

Pour la structure Dockerfile, nous recommandons d'utiliser un format similaire à l'exemple suivant :

```
FROM ghcr.io/astral-sh/uv:python3.13-bookworm-slim
WORKDIR /app

# All environment variables in one layer
ENV UV_SYSTEM_PYTHON=1 \
    UV_COMPILE_BYTECODE=1 \
    UV_NO_PROGRESS=1 \
    PYTHONUNBUFFERED=1 \
    DOCKER_CONTAINER=1 \
    AWS_REGION=us-east-1 \
```

```
    AWS_DEFAULT_REGION=us-east-1

COPY requirements.txt requirements.txt
# Install from requirements file
RUN uv pip install -r requirements.txt

RUN uv pip install aws-opentelemetry-distro>=0.10.1

# Signal that this is running in Docker for host binding logic
ENV DOCKER_CONTAINER=1

# Create non-root user
RUN useradd -m -u 1000 bedrock_agentcore
USER bedrock_agentcore

EXPOSE 9000
EXPOSE 8000
EXPOSE 8080

# Copy entire project (respecting .dockerignore)
COPY . .

# Use the full module path
CMD ["opentelemetry-instrument", "python", "-m", "server"]
```

Étape 2 : Testez votre serveur MCP localement

Avant le déploiement sur AWS, il est important de tester votre serveur MCP localement pour vous assurer qu'il fonctionne comme prévu. Pour obtenir des instructions détaillées sur les tests locaux, reportez-vous au manuel Amazon Bedrock AgentCore Developer Guide - Testez votre serveur MCP localement.

Étape 3 : Déploiement sur Amazon ECR

Une fois votre serveur MCP créé et testé localement, procédez comme suit pour le déployer sur Amazon ECR :

1. Assurez-vous que la dernière version de l'AWS CLI et de Docker est installée. Pour plus d'informations, consultez Getting Started with Amazon ECR.
2. Récupérez un jeton d'authentification et authentifiez votre client Docker auprès de votre registre. Utilisez la CLI AWS :

```
aws ecr get-login-password --region us-east-1 | docker login --username AWS --
password-stdin <account-id>.dkr.ecr.us-east-1.amazonaws.com
```

3. Créez votre image Docker à l'aide de la commande suivante. Pour plus d'informations sur la création d'un fichier Docker à partir de zéro, consultez la documentation Docker. Vous pouvez ignorer cette étape si votre image est déjà créée :

```
docker build -t <repository-name> .
```

4. Une fois la compilation terminée, balisez votre image afin de pouvoir la transférer vers ce référentiel :

```
docker tag <repository-name>:latest <account-id>.dkr.ecr.us-east-1.amazonaws.com/
<repository-name>:latest
```

5. Exécutez la commande suivante pour transférer cette image vers le dépôt AWS que vous venez de créer :

```
docker push <account-id>.dkr.ecr.us-east-1.amazonaws.com/<repository-name>:latest
```

Pour obtenir des instructions de déploiement complètes, reportez-vous au manuel Amazon Bedrock AgentCore Developer Guide - Déployez votre serveur MCP sur AWS.

Étape 4 : Utiliser l'URI ECR dans GAAB

Après avoir transféré avec succès votre image Docker vers Amazon ECR, copiez l'URI de l'image depuis la console ECR. Vous utiliserez cette URI lors du déploiement de votre serveur MCP via l'assistant de déploiement de Generative AI Application Builder sur AWS.

Étapes pour créer différentes cibles de passerelle MCP

Amazon Bedrock AgentCore Gateway vous permet de APIs transformer les services AWS existants en outils MCP utilisables par vos agents. La passerelle prend en charge plusieurs types de cibles, ce qui vous permet d'intégrer de manière fluide divers services principaux.

Les types de cibles suivants sont pris en charge :

- Objectifs Lambda : transformez les fonctions AWS Lambda en outils MCP. Pour obtenir des instructions détaillées, consultez le guide du AgentCore développeur Amazon Bedrock intitulé « Ajouter des cibles Lambda ».
- Cibles OpenAPI : utilisez les spécifications OpenAPI pour définir et exposer APIs REST en tant qu'outils MCP. Pour obtenir des instructions détaillées, reportez-vous au manuel Amazon Bedrock AgentCore Developer Guide - schéma OpenAPI.
- Objectifs de Smithy : créez des outils MCP à l'aide des définitions du modèle Smithy pour des intégrations d'API sécurisées. Pour obtenir des instructions détaillées, reportez-vous au manuel Amazon Bedrock AgentCore Developer Guide - Building Smithy targets.
- Cibles du serveur MCP : connectez-vous directement à des serveurs MCP externes via des points de terminaison URL, ce qui vous permet d'intégrer des serveurs MCP existants. Pour obtenir des instructions détaillées, reportez-vous au manuel Amazon Bedrock AgentCore Developer Guide - Les cibles des serveurs MCP.

Pour obtenir des exemples et des didacticiels supplémentaires sur la création de cibles MCP Gateway, consultez le référentiel d' AgentCore échantillons Amazon Bedrock.

Configuration d'une base de connaissances

Cette section explique comment intégrer des données dans la base de connaissances que vous avez sélectionnée pour la solution. La solution prend actuellement en charge les bases de connaissances Amazon Kendra et Amazon Bedrock en tant que bases de connaissances pour votre déploiement de cas d'utilisation basé sur RAG.

Amazon Kendra

Si vous utilisez Amazon Kendra comme base de connaissances, consultez le guide du développeur Amazon Kendra pour savoir comment utiliser les différents connecteurs de sources de données pour vous aider à ingérer des données provenant d'une large sélection de sources.

Important : pour éviter toute perte de données accidentelle, la solution ne supprime pas automatiquement l'index Kendra (qu'il ait été créé par la solution ou non) lors de la suppression d'un déploiement ou d'une pile. Si vous souhaitez supprimer votre base de connaissances et ne plus avoir à encourir de coûts, consultez la section sur la désinstallation manuelle pour savoir quelles ressources sont conservées et comment les nettoyer.

Bases de connaissances Amazon Bedrock

Les bases de connaissances Amazon Bedrock peuvent être soutenues par différents magasins vectoriels, chacun ayant la capacité d'indexer vos données. Pour configurer et enrichir votre base de connaissances, consultez le guide de l'utilisateur d'Amazon Bedrock. Plus précisément, vous souhaiterez :

- Configurez d'abord votre source de données
- Configurez ensuite un index vectoriel pour votre base de connaissances dans un magasin de vecteurs compatible. Notez que cela peut être ignoré si vous utilisez l'option « Création rapide d'un nouveau magasin de vecteurs » dans la console Bedrock lors de la création de la base de connaissances.
- Enfin, vous pouvez créer la base de connaissances et synchroniser vos sources de données configurées.

Paramètres avancés de la base de connaissances

Les paramètres avancés de la base de connaissances tels que le filtrage de la base de connaissances et le RAG avec contrôle d'accès basé sur les rôles peuvent être utilisés avec la solution. Le filtrage des bases de connaissances peut s'appliquer à l'une ou l'autre des bases de connaissances, tandis que RAG avec contrôle d'accès basé sur les rôles est spécifiquement disponible pour Amazon Kendra.

Filtrage de la base de connaissances

La solution vous permet de spécifier des filtres d'attributs Amazon Kendra ou des filtres de récupération de la base de connaissances Bedrock lors du déploiement d'un cas d'utilisation dans la section Configurations RAG avancées de l'étape de la base de connaissances de l'assistant. Ces filtres définissent la manière dont les sources de données de la base de connaissances sont interrogées, telles que les stratégies de recherche, les langues du document sous-jacent faisant l'objet des requêtes, etc.

Dans les deux cas, un objet JSON est utilisé pour spécifier les paramètres de filtre selon le format spécifié dans la documentation de chaque service (comme indiqué ci-dessus).

Exemple 1 : Kendra AttributeFilter

```
{
 "EqualsTo": {
```

```
 "Key": "_language_code",
 "Value": {
 "StringValue": "es"
 }
 }
}
```

Exemple 2 : Bedrock RetrievalFilter

```
{
 "equals": {
 "key": "language",
 "value": "es"
 }
}
```

RAG avec contrôle d'accès basé sur les rôles avec Amazon Kendra

Le contrôle d'accès basé sur les rôles (RBAC) permet de contrôler quels utilisateurs ou groupes peuvent accéder à certains documents de votre index Amazon Kendra ou voir certains documents dans leurs résultats de recherche. Pour configurer le RBAC pour votre identifiant d'index Amazon Kendra avec votre cas d'utilisation de Generative AI Application Builder on AWS (GAAB), procédez comme suit :

1. Configuration de l'index Amazon Kendra

1. Assurez-vous d'avoir créé un index Amazon Kendra et d'y avoir ajouté au moins une source de données.
2. Configurez le contrôle d'accès pour votre source de données en fonction des groupes d'utilisateurs. Pour une source de données S3, suivez les instructions disponibles dans la documentation pour configurer des listes de contrôle d'accès (ACLs) en utilisant les mêmes noms de groupe que ceux créés dans votre groupe d'utilisateurs Amazon Cognito. Cela garantit que les utilisateurs ne peuvent accéder qu'aux documents et aux résultats de recherche qu'ils sont autorisés à consulter en fonction de leur appartenance au groupe.

> **Note**
>
> Sous Contrôle d'accès utilisateur dans l'index Kendra que vous avez créé, laissez le contrôle d'accès utilisateur basé sur des jetons sur Non. Lorsque vous activez le contrôle

d'accès basé sur les rôles à l'étape 2, Generative AI Application Builder sur AWS extrait les demandes appropriées du jeton d'authentification utilisateur et crée un filtre d'attributs.

2. Déployer le cas d'utilisation de RAG à l'aide de l'assistant de déploiement GAAB

1. Suivez les instructions affichées à l'écran dans l'assistant de déploiement GAAB jusqu'à ce que vous atteigniez l'étape 4 de l'assistant pour configurer RAG.
2. À l'étape Sélectionner la base de connaissances de l'assistant de déploiement, choisissez Amazon Kendra comme type de base de connaissances.
3. Spécifiez si vous possédez déjà un index Amazon Kendra ou si vous souhaitez en créer un nouveau. Si vous possédez déjà un index, indiquez l'ID de votre index Amazon Kendra qui a été configuré avec des listes de contrôle d'accès (ACLs) basées sur des groupes d'utilisateurs.
4. Activez l'option Contrôle d'accès basé sur les rôles. Cette option garantit que les résultats de recherche renvoyés par l'index Amazon Kendra sont filtrés en fonction du rôle de l'utilisateur et des autorisations de groupe.
5. Passez en revue et déployez le cas d'utilisation.

3. Configurer Amazon Cognito

1. Localisez le groupe d'utilisateurs Amazon Cognito utilisé par votre déploiement GAAB. Ce groupe d'utilisateurs Amazon Cognito est généralement créé par le tableau de bord CloudFormation principal du déploiement.
2. Créez de nouveaux utilisateurs dans le groupe d'utilisateurs Amazon Cognito. Lorsque vous créez des utilisateurs, sélectionnez l'option « Envoyer une invitation par e-mail » afin que les utilisateurs reçoivent des informations de connexion temporaires par e-mail. Cela permet aux nouveaux utilisateurs de s'inscrire et d'accéder à l'application GAAB.
3. Créez des groupes d'utilisateurs dans le pool d'utilisateurs Amazon Cognito. Assurez-vous que les noms des groupes correspondent exactement aux groupes configurés dans votre index Amazon Kendra. ACLs Cela est crucial pour activer le RBAC, car l'appartenance au groupe de l'utilisateur déterminera les résultats de recherche auxquels il pourra accéder.
4. Affectez les utilisateurs aux groupes appropriés en fonction de leurs rôles et de leurs autorisations d'accès. Les utilisateurs doivent être ajoutés à la fois au groupe requis pour l'ACL de l'index Amazon Kendra et au groupe spécifique au cas d'utilisation créé lors du déploiement du GAAB.

Cela garantit que les utilisateurs disposent des autorisations nécessaires pour accéder au cas d'utilisation spécifique et aux résultats de recherche pertinents.

En suivant ces étapes, vous aurez configuré le contrôle d'accès basé sur les rôles (RBAC) pour votre déploiement GAAB, en veillant à ce que les utilisateurs puissent uniquement accéder et interagir avec les informations et fonctionnalités pour lesquelles ils sont autorisés, en fonction du groupe d'utilisateurs et des autorisations qui leur ont été attribués.

> **Note**
>
> À l'heure actuelle, seule Amazon Kendra prend en charge le RBAC pour les bases de connaissances dans le générateur d'applications d'IA générative sur AWS. Pour la base de connaissances Amazon Bedrock, le RBAC n'est pas pris en charge, mais vous pouvez utiliser des filtres de métadonnées pour atteindre un certain niveau de filtrage. Pour plus d'informations, consultez le guide de l'utilisateur d'Amazon Bedrock.

Configuration de vos invites

L'assistant du tableau de bord de déploiement comporte une étape de configuration rapide qui vous permet de personnaliser l'expérience instantanée et le modèle qui guidera les interactions entre les utilisateurs et le modèle d'IA. La configuration correcte de ces paramètres est essentielle pour obtenir des réponses précises et pertinentes de la part de l'assistant AI.

Cette section contrôle l'expérience globale et le comportement de l'invite AI.

- Longueur maximale du modèle d'invite : ce paramètre détermine la longueur maximale (en caractères) du modèle d'invite. Une valeur plus élevée permet de fournir plus de contexte au modèle d'IA, ce qui peut conduire à des réponses plus précises. Cependant, des instructions trop longues peuvent également générer du bruit et avoir un impact négatif sur les performances. Pour les modèles Amazon Bedrock, les valeurs par défaut pour la longueur maximale du modèle d'invite (en caractères) sont calculées à l'aide des limites de jetons du modèle sous-jacent. Si vous modifiez et changez le nom d'un modèle dans Bedrock, le bouton « Réinitialiser par défaut » est surligné et peut être utilisé pour adopter les paramètres par défaut du modèle nouvellement sélectionné. Pour les modèles Amazon SageMaker AI, des valeurs par défaut raisonnables sont fournies, mais il est recommandé de vérifier votre modèle sous-jacent et de choisir la longueur maximale du modèle d'invite et la longueur du texte de saisie en conséquence. Reportez-vous à la section Conseils sur la gestion des limites de jetons des modèles pour plus d'informations.

- Longueur maximale du texte saisi : ce paramètre limite la longueur maximale (en caractères) du texte saisi par l'utilisateur. Les entrées plus longues peuvent contenir des informations non pertinentes, ce qui augmente le risque d'obtenir des réponses non pertinentes ou inexactes à partir du modèle d'IA.
- Modification de l'invite utilisateur : cette option vous permet d'activer ou de désactiver la possibilité pour les utilisateurs de modifier le modèle d'invite via l'interface utilisateur du chat. La désactivation de cette fonctionnalité peut contribuer à maintenir la cohérence et à empêcher toute modification involontaire de l'invite.

Modèle d'invite

Cette section vous permet de définir le modèle d'invite réel qui sera utilisé par le modèle d'IA. Le modèle d'invite suit généralement une structure qui inclut des espaces réservés pour divers composants, tels que les entrées de l'utilisateur, les passages de référence et l'historique des discussions.

- Modèle d'invite : il s'agit de la zone de texte principale dans laquelle vous pouvez écrire ou coller le modèle d'invite souhaité. Le modèle doit être conçu pour fournir le contexte et les instructions nécessaires au modèle d'IA. Il inclut généralement les espaces réservés suivants :
  - `{input}`: Cet espace réservé est obligatoire pour les déploiements de Sagemaker AI et sera remplacé par la saisie ou la requête de l'utilisateur.
  - `{history}`: Cet espace réservé est obligatoire pour les déploiements de Sagemaker AI et sera remplacé par l'historique des discussions de la conversation en cours.
  - `{context}`: Cet espace réservé est obligatoire pour les déploiements RAG et sera remplacé par les extraits de documents obtenus à partir de la base de connaissances configurée.
- Reformuler la question ? : Cette option (disponible uniquement pour les déploiements RAG) détermine si la requête d'entrée initiale de l'utilisateur doit être reformulée ou désambiguïsée avant d'être transmise au modèle d'IA. La reformulation de la requête peut parfois aider le modèle à mieux comprendre l'intention de l'utilisateur, ce qui peut mener à des réponses plus précises.

Lors de la configuration du modèle d'invite et de l'expérience, il est essentiel de trouver un équilibre entre fournir suffisamment de contexte et d'instructions au modèle d'IA tout en évitant les informations trop longues ou non pertinentes susceptibles d'entraîner des problèmes de bruit ou de performance.

Paramètres d'invite avancés

Cette section vous permet de contrôler la façon dont l'historique des conversations est présenté au modèle d'IA.

- Taille de l'historique de suivi : ce paramètre détermine le nombre de messages précédents qui doivent être inclus dans l'invite finale. Si cette valeur est définie à zéro, aucun historique ne sera injecté dans le modèle d'invite ou dans le modèle d'invite de désambiguïsation. Remarque : même lorsqu'il est défini sur zéro, un espace réservé {history} doit toujours exister dans les modèles d'invite. Au moment de l'exécution, elle sera remplacée par une chaîne vide.
  - Remarque : Il est recommandé de fournir un nombre pair pour cette valeur. Si vous fournissez un nombre impair, seule la réponse de l'IA d'une interaction jumelée serait renvoyée.
- Préfixe humain : il s'agit du préfixe utilisé pour identifier les messages envoyés par l'utilisateur dans l'historique des conversations.
- Préfixe AI : il s'agit du préfixe utilisé pour identifier les messages renvoyés par le modèle AI dans l'historique des conversations.

Configuration de l'invite de désambiguïsation

Cette section vous permet de configurer le comportement et le modèle pour désambiguïser les entrées utilisateur avant de les envoyer à la base de connaissances configurée.

- Activer la désambiguïsation : cette option détermine si les entrées utilisateur doivent être désambiguïsées avant d'être envoyées à la base de connaissances.
- Modèle d'invite de désambiguïsation : il s'agit du modèle d'invite utilisé pour lever l'ambiguïté des saisies par les utilisateurs lorsqu'ils sont connectés à une base de connaissances. Le résultat généré à partir de cette invite sera utilisé comme requête envoyée à la base de connaissances. La désactivation de la désambiguïsation entraînerait l'envoi de la requête brute de l'utilisateur à la base de connaissances sans modification.

Par exemple, lorsque la désambiguïsation est activée, une demande de suivi de l'utilisateur intitulée « Combien ça coûte ? » pourrait être désambiguïsé en « Combien coûte le renouvellement de ma plaque d'immatriculation ? » , ce qui permet d'améliorer la requête de recherche.

Utilisation du scénario d'utilisation du texte déployé

L'interface utilisateur intégrée pour le cas d'utilisation du texte est destinée à permettre aux utilisateurs professionnels d'explorer et d'expérimenter rapidement le déploiement créé par

l'utilisateur administrateur. Les modifications de configuration effectuées par l'utilisateur professionnel ne prennent effet que pour sa session. L'utilisateur professionnel doit partager ces modifications avec l'utilisateur administrateur qui peut mettre à jour le déploiement de base avec ces modifications pour que tous puissent les utiliser.

L'interface utilisateur du chat comprend les éléments suivants :

- Fenêtre de discussion
- Zone de saisie du chat
- Settings
- Conversation claire

Fenêtre de discussion

Il prend différentes tournures de la conversation. Les messages commençant par la droite proviennent de l'utilisateur professionnel et ceux commençant par la gauche proviennent du LLM configuré. Une petite icône en forme de presse-papiers figure sur toutes les réponses LLM pour permettre de copier facilement les réponses.

Zone de saisie du chat

La zone de saisie du chat est épinglée au bas de la fenêtre de discussion. C'est ici que les utilisateurs professionnels peuvent saisir leurs messages à envoyer au LLM. Juste au-dessus de la zone de saisie se trouve l'état de la connexion. Si la connexion est perdue (par exemple, en raison de l'inactivité), une nouvelle connexion est automatiquement créée lors du prochain envoi d'un message de chat. Cette demande devrait prendre un peu plus de temps en raison du temps de WebSocket connexion supplémentaire.

En fonction de la configuration spécifique, une longueur maximale peut être imposée à l'entrée. Si cette limite est dépassée, les utilisateurs reçoivent une alerte et le message n'est pas envoyé.

Remarque : si vous utilisez RAG avec Amazon Kendra, l'API Retrieve tronquera les requêtes à 30 mots symboliques. Si vous vous attendez à de plus longues saisies par les utilisateurs, évaluez dans quelle mesure cela pourrait affecter les performances de recherche.

Settings

Pour permettre aux utilisateurs professionnels d'expérimenter rapidement différentes configurations, un panneau de paramètres est disponible, qui permet de on-the-fly modifier certaines options de configuration de déploiement

(exemple, modèle d'invite). Ces modifications ne peuvent être effectuées qu'au début d'une nouvelle session. Une fois qu'une conversation est lancée, l'effacement de la conversation permet de réactiver la modification des paramètres de configuration.

Remarque : Les utilisateurs administrateurs peuvent choisir de verrouiller les paramètres d'un déploiement. Ils peuvent empêcher les modifications en direct au moment du déploiement via l'assistant lors de l'étape d'invite.

Conversation claire

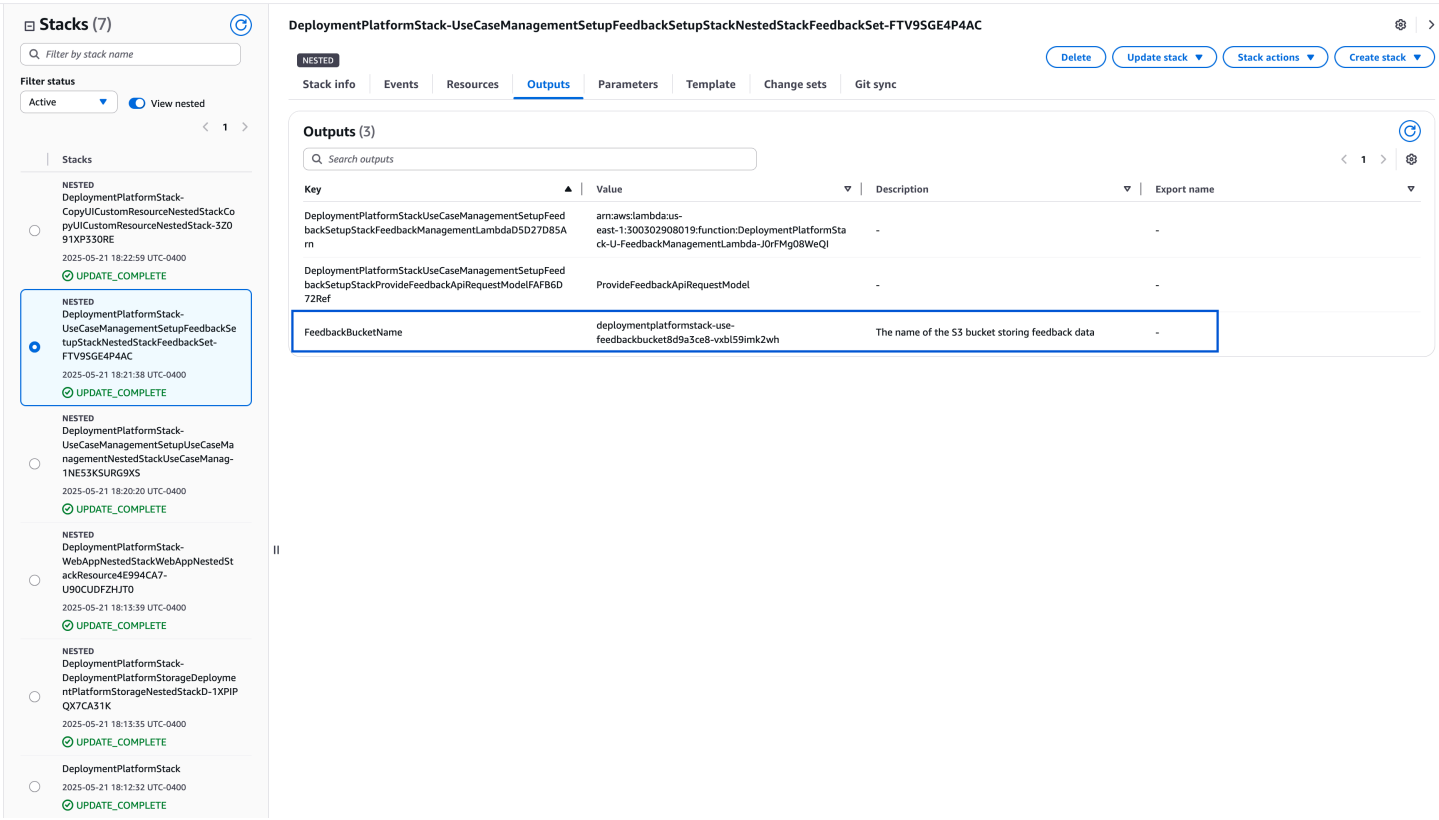
Au cours de la conversation, la solution conserve un historique des discussions, ce qui permet une expérience conversationnelle. Cela permet de désambiguïser les requêtes et de poser des questions de suivi. Pour réinitialiser une conversation et supprimer tout l'historique des discussions associées à cette interaction, choisissez *Effacer la conversation *en haut de la fenêtre de discussion. Une fois la conversation terminée, une nouvelle session est créée pour réactiver la modification des paramètres.

Accès et analyse des commentaires collectés par les utilisateurs

Depuis la version 3.0.0, le tableau de bord de déploiement déploie une pile de commentaires imbriquée qui permet aux cas d'utilisation des agents Text et Bedrock déployés avec le tableau de bord de disposer de la fonctionnalité de collecte de commentaires pour les réponses qu'il génère. LLM/Agent En particulier, les utilisateurs peuvent fournir un feedback positif ou négatif ainsi qu'un commentaire facultatif. Si l'utilisateur fournit un commentaire négatif, il peut ensuite sélectionner l'une des catégories négatives suivantes : « Inexact », « Incomplet ou insuffisant », « Nuisible », « Autre ». and/or

Une fois que l'utilisateur a fourni les commentaires, ceux-ci sont stockés dans un compartiment S3 partitionné par ID de cas d'utilisation, année et mois. L'ID de cas d'utilisation se trouve dans le tableau de bord de déploiement et le compartiment Feedback S3 se trouve dans les sorties de la pile imbriquée de commentaires de la pile du tableau de bord de déploiement :

Représente la pile de déploiement : recherche du nom du compartiment de commentaires



Les commentaires des utilisateurs sont envoyés sous forme de demande d'API contenant un ensemble minimal d'informations :

```
{
  "useCaseRecordKey": "a1b2c3d4-e5f6g7h8",
  "conversationId": "12345678-1234-1234-1234-123456789012",
  "messageId": "87654321-4321-4321-4321-210987654321",
  "rephrasedQuery": "What are the key features of the Generative AI Application Builder
 on AWS?",
  "sourceDocuments": [
    "s3://bucket-name/document1.pdf",
    "s3://bucket-name/document2.pdf"
  ],
  "feedback": "positive",
  "feedbackReason": [
    "Incomplete or insufficient"
  ],
  "comment": "The response was helpful but could include more details about important
 features."
}
```

Cette charge utile est ensuite traitée par un lambda à l'aide du `useCaseRecordKey` qui identifie la configuration correcte d'un cas d'utilisation au moment du déploiement. Cette configuration est utilisée pour obtenir des détails spécifiques sur le feedback, tels que le nom ConversationTable du et (contient toutes les conversations et les séquences de messages humains et IA), qui est ensuite utilisé pour récupérer le véritable `userInput` et`llmResponse`. Des détails supplémentaires sont également joints à cet enregistrement de commentaires, tels que le cas `agentId` et `agentAliasId` pour un cas d'utilisation de Bedrock Agent, et, etc. `modelProviderbedrockModelId`, pour un cas d'utilisation de type Text utilisant cette configuration. Pour plus de détails sur la façon d'accéder à cette configuration, consultez la section Mappages de commentaires personnalisés ci-dessous. Chaque demande de feedback entrante est stockée sous forme d'objet JSON et un exemple d'enregistrement de feedback peut ressembler à ceci pour un cas d'utilisation de type Text :


```
{
   "useCaseId": "12345678-1234-1234-1234-123456789012",
   "useCaseRecordKey": "c07a2e3b-2f31b1e0",
   "userId": "22345678-1234-1234-1234-123456789012",
   "conversationId": "dd51de5d-5af1-4ec6-91d2-aadf14352109",
   "messageId": "32345678-1234-1234-1234-123456789012",
   "userInput": "What are its key features?",
   "rephrasedQuery": "What are the key features of the Generative AI Application
 Builder on AWS?",
   "llmResponse": "Generative AI Application Builder on AWS can help you build
 production ready enterprise chatbots rapidly.",
   "feedback": "negative",
   "feedbackReason": [
      "Incomplete or insufficient"
   ],
   "comment": "The response was helpful but could include more details about important
 features.",
   "timestamp": "2025-05-22T18:48:08.340Z",
   "feedbackId": "42345678-1234-1234-1234-123456789012",
   "useCaseType": "Text",
   "modelProvider": "Bedrock",
   "bedrockModelId": "amazon.nova-lite-v1:0",
   "ragEnabled": "false"
}
```


ou comme ceci pour un cas d'utilisation de Bedrock Agent :


```
{
   "useCaseId": "12345678-1234-1234-1234-123456789012",
```

```
   "useCaseRecordKey": "c07a2e3b-2f31b1e0",
   "userId": "22345678-1234-1234-1234-123456789012",
   "conversationId": "dd51de5d-5af1-4ec6-91d2-aadf14352109",
   "messageId": "32345678-1234-1234-1234-123456789012",
   "userInput": "What are its key features?",
   "llmResponse": "Generative AI Application Builder on AWS can help you build
 production ready enterprise chatbots rapidly.",
   "feedback": "negative",
   "feedbackReason": [
      "Incomplete or insufficient"
   ],
   "comment": "The response was helpful but could include more details about important
 features.",
   "timestamp": "2025-05-22T18:48:08.340Z",
   "feedbackId": "42345678-1234-1234-1234-123456789012",
   "useCaseType": "Agent",
   "agentId": "AHFXUJCAK1",
   "agentAliasId": "KSEDKOS0BL"
}
```

Ce feedback peut ensuite être utilisé pour un traitement ultérieur, une analyse et une modélisation des boucles de réentraînement/de feedback. Vous pouvez également ajouter des mappages personnalisés pour améliorer l'enregistrement des commentaires stocké dans le lambda des commentaires.

Mappages de commentaires personnalisés

Le tableau de bord de déploiement contient un `LLMConfigTable` qui se trouve dans les sorties de la pile du tableau de bord de déploiement avec la clé`LLMConfigTableName`. `LLMConfigTable`contient les configurations pour chaque cas d'utilisation en fonction des paramètres sélectionnés par l'administrateur lors du déploiement du cas d'utilisation via l'assistant du tableau de bord de déploiement. Chaque configuration de cas d'utilisation est identifiée par son. `useCaseRecordKey` Voici un exemple d'enregistrement de configuration de cas d'utilisation dans le : `LLMConfigTable`

```
{
   "key": "2dd76cfa-bc1a14da",
   "config": {
      "ConversationMemoryParams": {
        ...
      },
```

```
      "FeedbackParams": {
         "CustomMappings": {
            "NumberOfDocs": "$.KnowledgeBaseParams.NumberOfDocs",
            "ScoreThreshold": "$.KnowledgeBaseParams.ScoreThreshold"
         },
         "FeedbackEnabled": true
      },
      "IsInternalUser": "true",
      "KnowledgeBaseParams": {
         "KendraKnowledgeBaseParams": {
            "ExistingKendraIndexId": "d2831033-667f-4539-ab28-e6c7c7c5988b",
            "RoleBasedAccessControlEnabled": false
         },
         "KnowledgeBaseType": "Kendra",
         "NumberOfDocs": 5,
         "ReturnSourceDocs": false,
         "ScoreThreshold": 0.3
      },
      "LlmParams": {
         "BedrockLlmParams": {
            "BedrockInferenceType": "QUICK_START",
            "ModelId": "amazon.nova-lite-v1:0"
         },
         "ModelParams": {},
         "ModelProvider": "Bedrock",
         "PromptParams": {
            ...
         },
         "RAGEnabled": true,
         "Streaming": false,
         "Temperature": 0.1,
         "Verbose": false
      },
      "UseCaseName": "test-rag-usecase",
      "UseCaseType": "Text"
   }
}
```

Si le feedback est activé pour un cas d'utilisation, cette configuration contiendra un `FeedbackParams` objet dans lequel un `CustomMappings` objet peut spécifier tous les champs supplémentaires JSONPaths à ajouter à l'enregistrement JSON de feedback stocké dans le compartiment de feedback S3. Par exemple, pour l'exemple de configuration de cas d'utilisation ci-dessus, le CustomMappings contient `NumberOfDocs` et `ScoreThreshold` JSONPaths en

plus dans l'`CustomMappings`objet qui commencent `config` par la racine du. JSONPath Avec cette configuration, chaque enregistrement JSON stocké dans le compartiment S3 de feedback commencera à obtenir ces 2 valeurs supplémentaires en plus des champs déjà fournis.

Analyse des données de feedback

Les données de feedback sont stockées dans S3 sous forme d'objets JSON. Voici quelques approches pour rendre ces données de feedback plus accessibles et exploitables :

Utilisation d'AWS Glue et d'Amazon Athena

AWS Glue et Amazon Athena fournissent un moyen sans serveur de cataloguer, d'interroger et d'analyser les données de vos commentaires.

AWS Glue vous permet de créer un robot d'exploration AWS Glue qui inspecte les données d'un compartiment S3, en déduit le schéma et enregistre toutes les métadonnées pertinentes dans un catalogue. Après cela, des services tels qu'Amazon Athena peuvent être utilisés pour interroger les données.

Vous pouvez consulter la documentation AWS Athena pour connaître les étapes à suivre pour connecter le compartiment S3 de commentaires à Amazon Athena à l'aide d'AWS Glue Data Catalog. Vous pouvez également utiliser certaines des fonctionnalités les plus puissantes de Glue pour effectuer des tâches d'extraction, de transformation et de chargement (ETL) sur ces données et les transformer dans un format adapté à vos analyses ou à vos cas d'utilisation de reconversion de modèles. Avec Glue, vous pouvez effectuer des opérations telles que le filtrage des enregistrements avec certains types de commentaires, le remplissage des informations manquantes, et vous pouvez également charger ces données dans un autre emplacement de stockage tel qu'un autre compartiment S3 ou un autre magasin de données AWS.

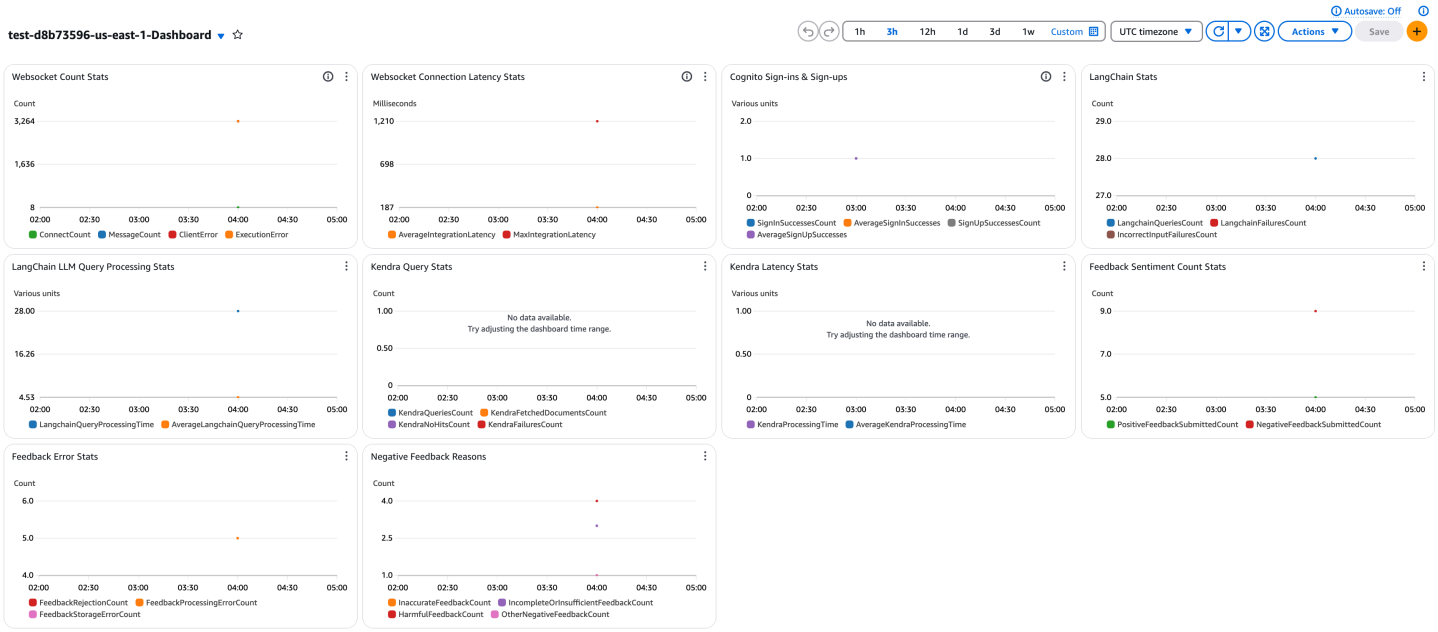
> **Note**
>
> En fonction de votre cas d'utilisation, pensez à programmer le crawler Glue pour qu'il s'exécute périodiquement (par exemple, chaque semaine) plutôt que tous les soirs afin d'optimiser les coûts, car les données de feedback peuvent être rares.

Utilisation des tableaux de CloudWatch bord de la solution

Vous avez également accès à un CloudWatch tableau de bord fourni avec la solution qui peut vous fournir des tendances en matière de commentaires positifs et négatifs, de catégories de raisons de

commentaires négatifs, etc., par cas d'utilisation. Vous pouvez trouver ce tableau de bord sous le nom de votre cas d'utilisation dans la section Tableaux de bord de la console AWS : CloudWatch

Représente le tableau de bord Usecase CloudWatch



Vous pouvez également créer des widgets supplémentaires dans ce tableau de bord ou créer des tableaux de bord Amazon Quick Sight.

Meilleures pratiques pour l'analyse des données de feedback

- Mettez en œuvre des politiques de cycle de vie des données dans votre compartiment S3 afin d'archiver les anciennes données de feedback vers des niveaux de stockage moins coûteux
- Créez une analyse distincte pour chaque cas d'utilisation afin d'identifier les opportunités d'amélioration spécifiques au modèle
- Établissez des seuils de feedback qui déclenchent des alertes lorsque les commentaires négatifs dépassent les niveaux acceptables
- Exportez régulièrement des informations critiques pour les partager avec les parties prenantes et les équipes d'amélioration des modèles

Afficher les métriques opérationnelles pour un déploiement

Le tableau de bord de déploiement et les piles de cas d'utilisation sont chacun dotés de leur propre CloudWatch tableau de bord permettant de suivre les différentes mesures opérationnelles de

la solution. Vous pouvez utiliser ces CloudWatch tableaux de bord pour comparer les différents déploiements. Pour accéder aux tableaux de bord :

1. Accédez à la console CloudWatch .
2. Recherchez les tableaux de bord prédéfinis en recherchant le nom de la pile ou l'identifiant unique universel (UUID).

Par exemple, le cas d'utilisation du texte est fourni avec des graphiques permettant de suivre le nombre de WebSocket connexions, le nombre de connexions et d'inscriptions d'utilisateurs, le temps nécessaire au LLM pour traiter une finalisation, etc. Les clients peuvent utiliser ces graphiques pour comparer différentes métriques _quantitatives d'un déploiement.

Example

Il est difficile de comparer les résultats qualitatifs des différents modèles appliqués à différents cas d'utilisation. Utilisez la fonction Clone pour lancer rapidement plusieurs déploiements afin de pouvoir comparer les résultats côte à côte.

Informations sur CloudWatch les journaux d'accès

Cette solution enregistre les messages d'erreur, d'avertissement, d'information et de débogage pour les fonctions Lambda. Pour choisir le type de messages à enregistrer :

1. Recherchez la fonction applicable dans la console AWS Lambda.
2. Ajoutez une variable d'environnement POWERTOOLS_LOG_LEVEL.
3. Définissez la variable sur le type de message applicable.

Pour obtenir des instructions supplémentaires, consultez la section Création de variables d'environnement Lambda dans le guide du développeur AWS Lambda.

Le tableau suivant répertorie les types de niveaux de journalisation parmi lesquels vous pouvez choisir.

Niveau	Description
ERREUR	Les journaux contiennent des informations sur tout ce qui entraîne l'échec d'une opération.

Niveau	Description
WARNING	Les journaux contiennent des informations sur tout élément susceptible de provoquer des incohérences dans la fonction, mais pas nécessairement de provoquer l'échec de l'opération. Les journaux incluent également des messages d'erreur.
INFOS	Les journaux contiennent des informations de haut niveau sur le fonctionnement de la fonction. Les journaux contiennent également des messages d'erreur et d'avertissement.
DÉBOGAGE	Les journaux contiennent des informations qui peuvent être utiles lors du débogage d'un problème lié à la fonction. Les journaux incluent également des messages d'erreur, d'avertis sement et d'information.

Utilisez la procédure suivante pour ajouter CloudWatch Logs Insights à cette solution.

1. Identifiez les groupes de journaux pertinents :
   a. Connectez-vous à la CloudFormation console AWS.
   b. Choisissez votre pile cible.
   c. Sélectionnez l'onglet Ressources et recherchez vos fonctions Lambda cibles.
   d. Connectez-vous à la console AWS Lambda et choisissez chacune de vos fonctions Lambda cibles.
   e. Pour chacune de vos fonctions Lambda cibles, sélectionnez l'onglet Monitor et choisissez View CloudWatch Logs.
   f. Copiez les noms des groupes de journaux dont vous souhaitez extraire des informations.
2. Accédez à la CloudWatch console Amazon.
3. Dans le menu de navigation, sous Logs, sélectionnez Logs Insights.
4. Sur la page Logs Insights, choisissez l'onglet Logs.
5. Recherchez les noms des groupes de journaux à partir de l'étape 1.

6. Copiez l'un des exemples de requêtes suivants et collez-le dans le champ de requête :

   a. Pour identifier toutes les exceptions du client :

   ```
   fields @message
   |filter @message like /(?i)Exception/|stats count(*) as exceptionCount by @message
   ```

   b. Pour récupérer le nombre d'appels par nom de fonction :

   ```
   stats count(*) by function_name
   ```

   c. Pour récupérer le nombre d'invocations sur des intervalles de cinq minutes, procédez comme suit :

   ```
   stats count(*) as invocations by bin(5m)
   ```

   d. Pour récupérer toutes les traces d'AWS X-Ray, procédez comme IDs suit :

   ```
   filter @message like "XRAY TraceId"
   |parse @message "XRAY TraceId: * " as traceId|stats count(*) by traceId
   ```

   e. Pour récupérer les journaux relatifs à un X-Ray Trace ID spécifique, procédez comme suit :

   ```
   filter @message like "your-traceid-here"
   ```

   f. Pour récupérer des WebSocket erreurs non autorisées :

   ```
   fields
   @ingestionTime,
   @log,
   @logStream,
   @message,
   @requestId,
   @timestamp,
   errorMessage,
   errorType
   |filter @message like /Unauthorized/ and @message like /websocket/|sort @timestamp
    desc
   ```

   g. Pour récupérer le nombre de métriques publiées :

   ```
   filter @message like "CloudWatchMetrics"
   ```

```
|parse @message /"Metrics":\s*\[(?<metrics>.*?)\]/|stats count(*) as metric_count
 by metrics
```

Guide du développeur

Cette section fournit le code source de la solution, un guide d'intégration, un guide de personnalisation et une référence d'API.

Code source

Consultez notre GitHub référentiel pour télécharger les fichiers source de cette solution et partager vos personnalisations avec d'autres utilisateurs.

Les modèles Generative AI Application Builder sur AWS sont générés à l'aide du AWS Cloud Development Kit (AWS CDK). Consultez le fichier README.md pour plus d'informations.

Guide d'intégration

L'ensemble de la solution est conçu pour être facilement extensible. La couche d'orchestration de cette solution est construite à l'aide LangChainde. Vous pouvez ajouter à cette solution n'importe quel fournisseur de modèles, base de connaissances ou type de mémoire de conversation pris en charge par LangChain (ou par un tiers fournissant des LangChain connecteurs pour ces composants).

Extension prise en charge LLMs

Pour ajouter un autre fournisseur de modèles, tel qu'un fournisseur LLM personnalisé, vous devez mettre à jour les trois composants suivants de la solution :

1. Créez une nouvelle pile `TextUseCase` CDK, qui déploie l'application de chat configurée avec votre fournisseur LLM personnalisé :
   a. Clonez le GitHub référentiel de cette solution et configurez votre environnement de génération en suivant les instructions fournies dans le fichier README.md.
   b. Copiez (ou créez-en un nouveau), collez-le dans le même répertoire et renommez-le `encustom-chat-stack.ts`. `source/infrastructure/lib/bedrock-chat-stack.ts`
   c. Renommez la classe du fichier en une classe appropriée, telle que`CustomLLMChat`.
   d. Vous pouvez choisir d'ajouter un secret Secrets Manager à cette pile, qui stocke vos informations d'identification pour votre LLM personnalisé. Vous pouvez récupérer ces informations d'identification lors de l'invocation du modèle dans la couche Lambda du chat décrite dans le paragraphe suivant.

2. Créez et attachez une couche Lambda contenant la bibliothèque Python du fournisseur de modèles à ajouter. Pour une application de chat utilisant Amazon Bedrock, la bibliothèque `langchain-aws` Python contient les connecteurs personnalisés situés au-dessus LangChain du package pour se connecter aux fournisseurs de modèles AWS (Amazon Bedrock et SageMaker AI), aux bases de connaissances (Amazon Kendra et Amazon Bedrock Knowledge Bases) et aux types de mémoire (tels que DynamoDB). De même, les autres fournisseurs de modèles disposent de leurs propres connecteurs. Cette couche vous permet d'associer la bibliothèque Python de ce fournisseur de modèles afin que vous puissiez utiliser ces connecteurs dans la couche Lambda du chat, qui invoque le LLM (étape 3). Dans cette solution, un bundler d'actifs personnalisé est utilisé pour créer des couches Lambda, qui sont attachées à l'aide des aspects CDK. Pour créer une nouvelle couche pour la bibliothèque de fournisseurs de modèles personnalisés, procédez comme suit :

   a. Accédez à la `LambdaAspects` classe dans le `source/infrastructure/lib/utils/lambda-aspects.ts` fichier.

   b. Suivez les instructions pour étendre les fonctionnalités de la classe d'aspects Lambda fournie dans le fichier (par exemple en ajoutant la `getOrCreateLangchainLayer` méthode). Pour utiliser cette nouvelle méthode (par exemple,`getOrCreateCustomLLMLayer`), mettez également à jour l'`LLM_LIBRARY_LAYER_TYPE`Sénumération dans le `source/infrastructure/lib/utils/constants.ts` fichier.

3. Étendez la fonction `chat` Lambda pour implémenter un générateur, un client et un gestionnaire pour le nouveau fournisseur.

   `source/lambda/chat`Il contient les LangChain connexions pour les différentes classes LLMs ainsi que les classes de support pour les créer LLMs. Ces classes de support suivent les modèles de conception orientés Builder et Object pour créer le LLM.

   Chaque gestionnaire (par exemple`bedrock_handler.py`) crée d'abord un client, vérifie les variables d'environnement requises dans l'environnement, puis appelle une `get_model` méthode pour obtenir la classe LangChain LLM. La méthode generate est ensuite appelée pour invoquer le LLM et obtenir sa réponse. LangChain prend actuellement en charge la fonctionnalité de streaming pour Amazon Bedrock, mais pas l' SageMaker IA. Sur la base d'une fonctionnalité de streaming ou non, le WebSocket gestionnaire approprié (`WebsocketStreamingCallbackHandler`ou`WebsocketHandler`) est appelé pour renvoyer la réponse à la WebSocket connexion à l'aide de la `post_to_connection` méthode.

   Le `clients/builder` dossier contient les classes qui aident à créer un générateur LLM à l'aide du modèle Builder. Tout d'abord, a `use_case_config` est extrait d'un magasin de configurations

DynamoDB, qui contient les informations relatives au type de base de connaissances, de mémoire de conversation et de modèle à construire. Il contient également les détails pertinents du modèle, tels que les paramètres du modèle et les instructions. Le générateur aide ensuite à suivre les étapes de création d'une base de connaissances, de création d'une mémoire de conversation pour maintenir le contexte de conversation pour le LLM, de définition des LangChain rappels appropriés pour les cas de streaming et de non-streaming, et de création d'un modèle LLM basé sur les configurations de modèle fournies. La configuration DynamoDB est stockée au moment de la création des cas d'utilisation lorsque vous déployez un cas d'utilisation à partir du tableau de bord de déploiement (ou lorsqu'il est fourni par les utilisateurs dans le cadre de déploiements de pile de cas d'utilisation autonomes sans le tableau de bord de déploiement).

Le `clients/factories` sous-dossier permet de définir la mémoire de conversation et la classe de base de connaissances appropriées, en fonction de la configuration LLM. Cela permet une extension facile à toute autre base de connaissances ou à tout autre type de mémoire que vous souhaitez que votre implémentation prenne en charge.

Le `shared` sous-dossier contient des implémentations spécifiques de la base de connaissances et de la mémoire de conversation qui sont instanciées dans les usines par le constructeur. Il contient également des récupérateurs de la base de connaissances Amazon Kendra et Amazon Bedrock appelés pour récupérer des documents relatifs LangChain aux cas d'utilisation du RAG, ainsi que des rappels, utilisés par le modèle LLM. LangChain

Les LangChain implémentations utilisent le langage LangChain d'expression (LCEL) pour composer ensemble des chaînes de conversation. `RunnableWithMessageHistory`La classe est utilisée pour conserver l'historique des conversations avec des chaînes LCEL personnalisées, permettant des fonctionnalités telles que le renvoi de documents sources et l'utilisation de la question reformulée (ou désambiguïsée) envoyée à la base de connaissances pour être également envoyée au LLM.

Pour créer votre propre implémentation d'un fournisseur personnalisé, vous pouvez :

a. Copiez le `bedrock_handler.py` fichier et créez votre gestionnaire personnalisé (par exemple,`custom_handler.py`), qui crée votre client personnalisé (par exemple,`CustomProviderClient`) (spécifié à l'étape suivante).
b. Copiez `bedrock_client.py` dans le dossier des clients. Renommez-le en `custom_provider_client.py` (ou renommez-le en votre nom de fournisseur de modèles spécifique, par exemple`CustomProvider`). Nommez la classe qu'elle contient de manière appropriée, par exemple `CustomProviderClient` qui hérite`LLMChatClient`.

Vous pouvez utiliser les méthodes fournies par `LLMChatClient` ou écrire vos propres implémentations pour les remplacer.

La `get_model` méthode crée un `CustomProviderBuilder` (voir l'étape suivante) et appelle la `construct_chat_model` méthode qui construit le modèle de chat à l'aide des étapes du générateur. Cette méthode agit en tant que directeur dans le modèle de générateur.

c. Copiez-le `clients/builders/bedrock_builder.py` et renommez-le en `custom_provider_builder.py` ainsi `CustomProviderBuilder` que la classe qu'il contient en héritant LLMBuilder ()`llm_builder.py`. Vous pouvez utiliser les méthodes fournies par LLMBuilder ou écrire vos propres implémentations pour les remplacer. Les étapes du générateur sont appelées en séquence dans la `construct_chat_model` méthode du client, telles que `set_model_defaultsset_knowledge_base`, et`set_conversation_memory`.

La `set_llm_model` méthode créerait le modèle LLM réel en utilisant toutes les valeurs définies à l'aide des méthodes appelées auparavant. Plus précisément, vous pouvez créer un LLM RAG (`CustomProviderRetrievalLLM`) ou non RAG (`CustomProviderLLM`), sur la base de `rag_enabled variable` ce qui est extrait de la configuration LLM dans DynamoDB.

Cette configuration est récupérée dans la `retrieve_use_case_config` méthode de la `LLMChatClient` classe.

d. Implémentez votre `CustomProviderRetrievalLLM` implémentation `CustomProviderLLM` ou son implémentation dans le `llm_models` sous-dossier selon que vous avez besoin d'un cas d'utilisation RAG ou non. La plupart des fonctionnalités permettant d'implémenter ces modèles sont fournies dans leurs `RetrievalLLM` classes `BaseLangChainModel` et respectivement, pour les cas d'utilisation autres que RAG et RAG.

Vous pouvez copier le `llm_models/bedrock.py` fichier et apporter les modifications nécessaires pour appeler le LangChain modèle qui fait référence à votre fournisseur personnalisé. Par exemple, Amazon Bedrock utilise une `ChatBedrock` classe pour créer un modèle de chat en utilisant LangChain.

La méthode generate génère la réponse LLM à l'aide des chaînes LangChain LCEL.

Vous pouvez également utiliser `get_clean_model_params` cette méthode pour nettoyer les paramètres du modèle conformément LangChain aux exigences de votre modèle.

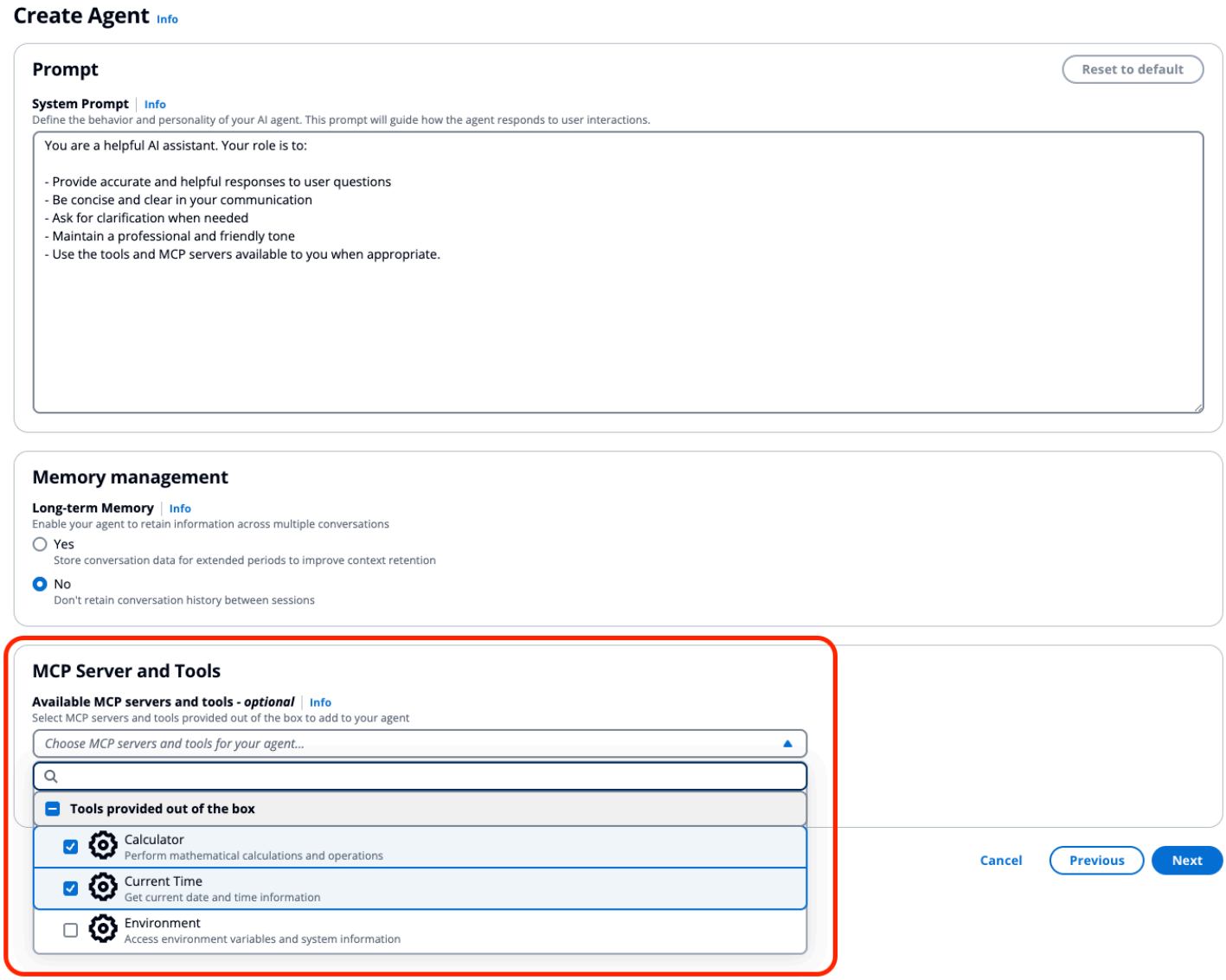
Extension des outils Strands pris en charge

La solution vous permet de créer et de déployer des serveurs MCP, des agents d'intelligence artificielle et des flux de travail multi-agents. Dans le cadre de l'expérience Agent Builder, vous pouvez associer des serveurs MCP pour offrir à vos agents des fonctionnalités supplémentaires. Outre les serveurs MCP, vous pouvez tirer parti des outils intégrés fournis par Strands (le framework sous-jacent utilisé par la solution).

Prête à l'emploi, la solution est préconfigurée avec les outils Strands suivants :

- Heure actuelle (activée par défaut)
- Calculatrice (activée par défaut)
- Environnement

Sélection du serveur et des outils MCP dans l'assistant Agent Builder affichant les outils Strands intégrés



Pour étendre à vos agents des outils Strands supplémentaires, suivez le processus en quatre étapes décrit dans cette section.

Étape 1 : Trouvez l'outil Strands

Parcourez les outils Strands disponibles pour identifier l'outil que vous souhaitez utiliser. Chaque outil possède des capacités et des exigences de configuration spécifiques.

Par exemple, pour ajouter les fonctionnalités de récupération de la base de connaissances Amazon Bedrock, vous devez utiliser l'outil de récupération.

Étape 2 : mise à jour du paramètre SSM

Pour rendre un outil disponible dans l'interface utilisateur de déploiement d'Agent Builder, mettez à jour le paramètre AWS Systems Manager Parameter Store qui définit les outils Strands pris en charge.

1. Accédez au AWS Systems Manager Parameter Store dans votre compte AWS.
2. Localisez le paramètre : `/gaab/<stack-name>/strands-tools`
3. Ajoutez la configuration de votre outil à la fin de la liste existante à l'aide de la structure JSON suivante :

```
{
  "name": "Bedrock KB Retrieve",
  "description": "Retrieve information from Bedrock Knowledge Base",
  "value": "retrieve",
  "category": "AI",
  "isDefault": false
}
```

Champ	Description
nom	Nom d'affichage affiché dans l'interface utilisateur d'Agent Builder
description	Brève description des fonctionnalités de l'outil
valeur	Le nom exact de l'outil tel que défini dans le package d'outils Strands
category	Catégorie organisationnelle pour le regroupement des outils dans l'interface utilisateur
est par défaut	Si l'outil doit être activé par défaut pour les nouveaux agents

Étape 3 : Configuration des variables d'environnement

De nombreux outils Strands nécessitent des variables d'environnement pour la configuration. Vous pouvez définir ces variables de deux manières :

Option 1 : Configuration directe sur AgentCore Runtime

Mettez à jour l'agent déployé directement sur Amazon Bedrock AgentCore Runtime avec les variables d'environnement requises.

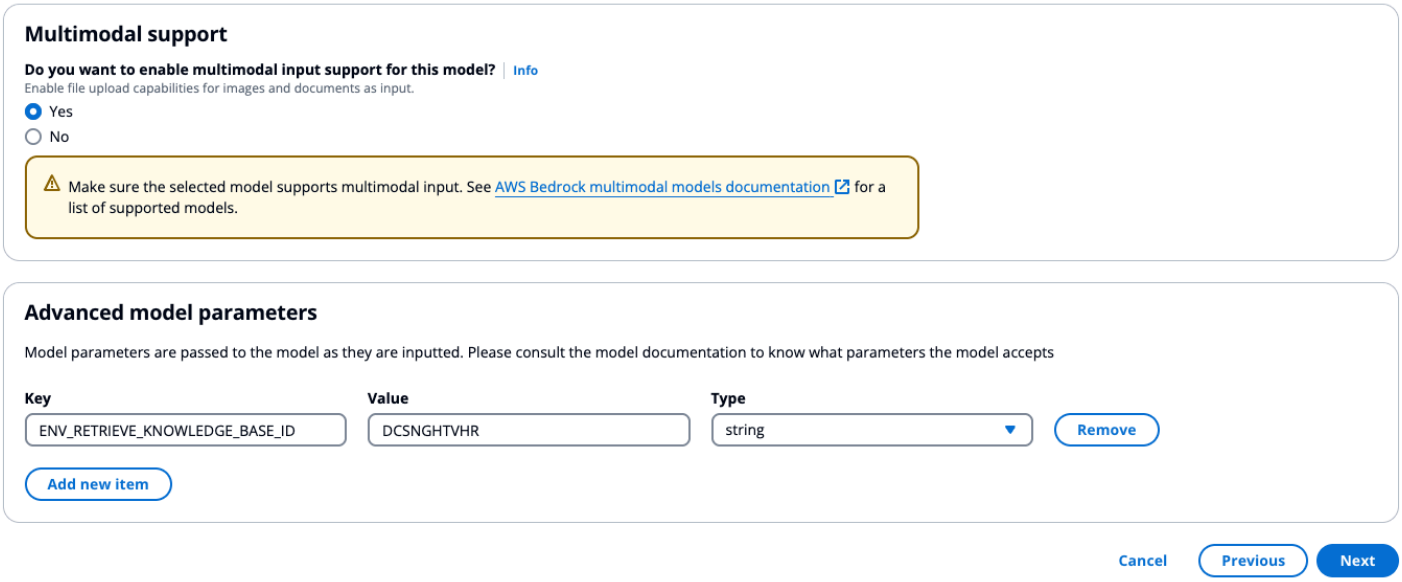
Option 2 : paramètres du modèle dans l'assistant de déploiement

Ajoutez des variables d'environnement lors de l'étape de sélection du modèle dans l'assistant Agent Builder à l'aide de la section Paramètres du modèle. Les variables d'environnement qui suivent la convention de dénomination `ENV_<ALL_CAPS_TOOL_NAME>_<env_variable_name>` seront automatiquement chargées au moment de l'exécution dans l'environnement d'exécution de l'agent en tant que`<env_variable_name>`.

Par exemple :

- `ENV_RETRIEVE_KNOWLEDGE_BASE_ID`devient `KNOWLEDGE_BASE_ID`
- `ENV_RETRIEVE_MIN_SCORE`devient `MIN_SCORE`

Section des paramètres avancés du modèle montrant la configuration ENV_RETRIEVE_KNOWLEDGE_BASE_ID



Reportez-vous à la documentation ou au code source de l'outil spécifique pour identifier les variables d'environnement requises. Pour l'outil de récupération, vous trouverez les options de configuration dans le code source.

Étape 4 : Ajouter des autorisations IAM

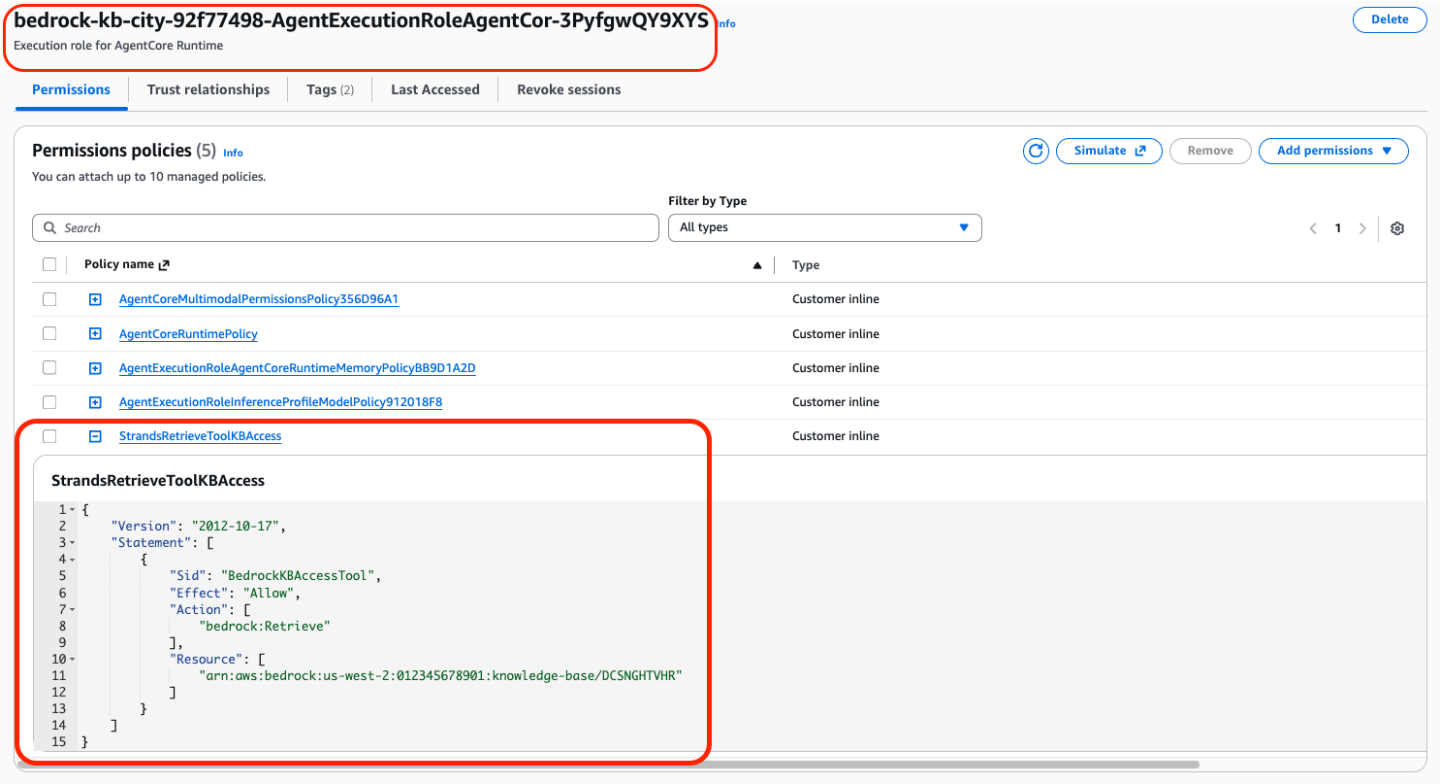
Ajoutez manuellement les autorisations IAM nécessaires à votre rôle AgentCore d'exécution Runtime pour permettre à l'agent d'utiliser l'outil.

Par exemple, pour utiliser l'outil de récupération avec les bases de connaissances Amazon Bedrock :

1. Accédez à la console IAM dans votre compte AWS.
2. Localisez le rôle AgentCore d'exécution Runtime pour votre agent.
3. Ajoutez l'autorisation suivante :

```
{
  "Effect": "Allow",
  "Action": "bedrock:Retrieve",
  "Resource": "arn:aws:bedrock:region:account-id:knowledge-base/knowledge-base-id"
}
```

Console IAM affichant la StrandsRetrieveTool KBAccess politique associée au rôle d' AgentCore exécution Runtime

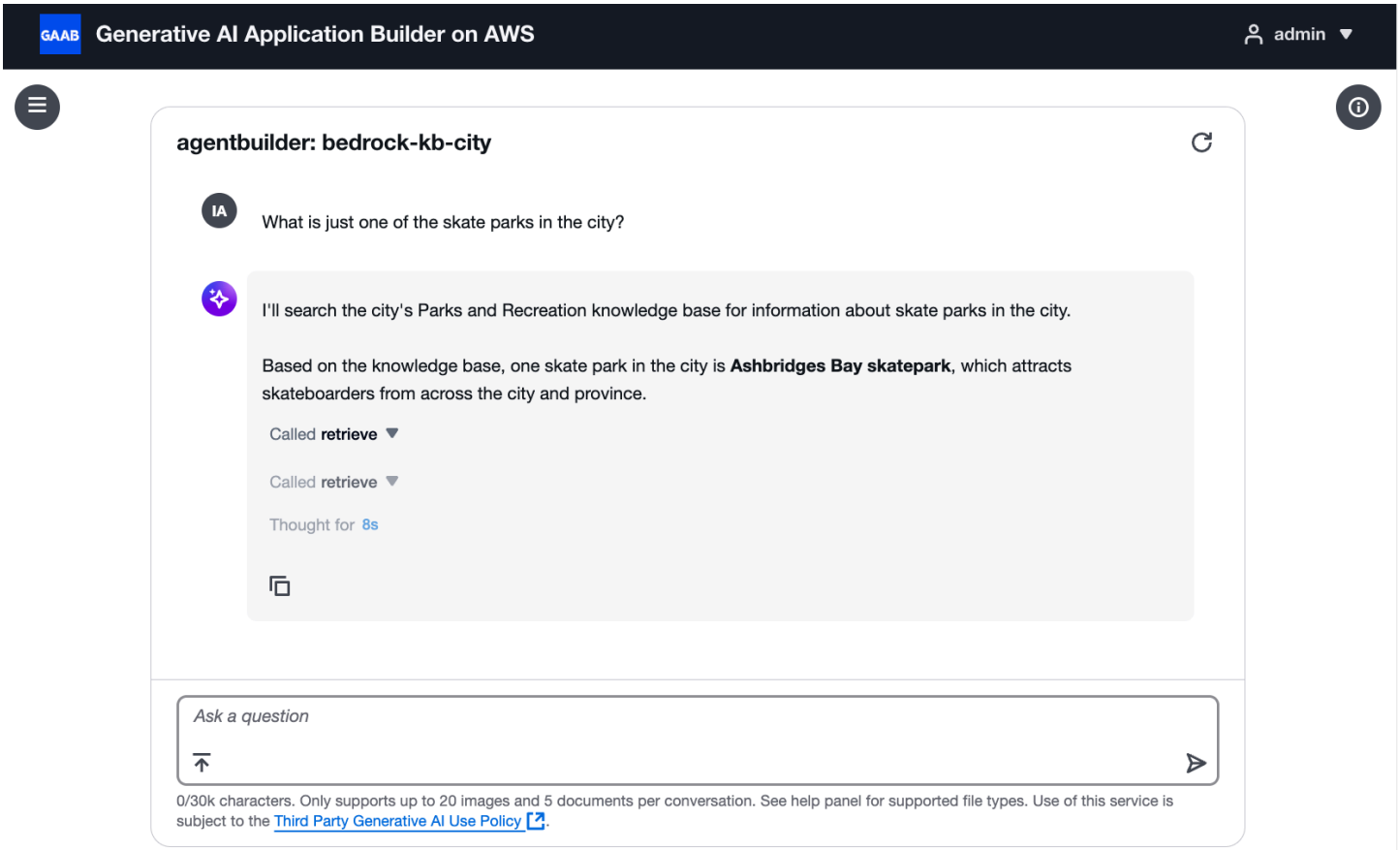


Les autorisations spécifiques requises varient en fonction de l'outil. Consultez la documentation de l'outil et la documentation du service AWS pour déterminer les autorisations IAM appropriées.

Étape 5 : tester l'agent

Une fois les étapes de configuration terminées, testez votre agent pour vérifier que l'outil fonctionne correctement. Vous devriez voir les appels d'outils dans les journaux d'exécution et les réponses de l'agent.

L'agent a utilisé avec succès l'outil de récupération pour répondre à une question sur les skate parks



> **Note**
>
> Pour une liste complète des outils Strands disponibles et de leurs fonctionnalités, reportez-vous à la documentation des outils communautaires Strands.

Élargir les bases de connaissances et les types de mémoires de conversation pris en charge

Pour ajouter vos implémentations de mémoire de conversation ou de base de connaissances, ajoutez les implémentations requises dans le `shared` dossier, puis modifiez les usines et les énumérations appropriées pour créer une instance de ces classes.

Lorsque vous fournissez la configuration LLM, qui est stockée dans le magasin de paramètres, la mémoire de conversation et la base de connaissances appropriées sont créées pour votre LLM. Par exemple, lorsque le `ConversationMemoryType` est spécifié en tant que DynamoDB, une instance `DynamoDBChatMessageHistory` de (disponible `shared_components/memory/`

ddb_enhanced_message_history.py à l'intérieur) est créée. Lorsque le KnowledgeBaseType est spécifié comme Amazon Kendra, une instance de KendraKnowledgeBase (disponible à l'intérieurshared_components/knowledge/kendra_knowledge_base.py) est créée.

Création et déploiement des modifications du code

Créez le programme à l'aide de la npm run build commande. Une fois les erreurs résolues, cdk synth lancez-vous pour générer les fichiers modèles et tous les actifs Lambda.

1. Vous pouvez utiliser le 0/stage-assets.sh script pour transférer manuellement tous les actifs générés vers le bucket intermédiaire de votre compte.
2. Utilisez la commande suivante pour déployer ou mettre à jour la plateforme :

   ```
   cdk deploy DeploymentPlatformStack --parameters AdminUserEmail='admin-
   email@amazon.com'
   ```

   Tous CloudFormation les paramètres AWS supplémentaires doivent également être fournis avec le AdminUserEmailparamètre.

Guide de personnalisation

Gestion du groupe d'utilisateurs de Cognito

Lorsque le tableau de bord de déploiement est déployé, un groupe d'utilisateurs Amazon Cognito ainsi qu'un utilisateur administrateur sont créés pour authentifier l'application. Ce groupe d'utilisateurs est partagé entre le tableau de bord de déploiement et tous les cas d'utilisation. L'utilisateur administrateur créé lors du déploiement du tableau de bord est automatiquement autorisé à accéder à tous les cas d'utilisation déployés à l'aide du tableau de bord. Ce mécanisme est fourni par le biais de groupes d'utilisateurs Amazon Cognito.

Lorsqu'un cas d'utilisation est déployé depuis le tableau de bord, si un e-mail est fourni, un utilisateur est créé dans le groupe d'utilisateurs partagé, ainsi qu'un groupe d'utilisateurs nommé pour le cas d'utilisation spécifique. L'utilisateur nouvellement créé est ensuite ajouté au groupe, ce qui lui permet d'accéder au cas d'utilisation.

Si vous souhaitez ajouter un utilisateur supplémentaire à un cas d'utilisation donné, vous pouvez le faire en créant un utilisateur dans le groupe d'utilisateurs de Cognito et en l'ajoutant au ou aux groupes correspondant aux cas d'utilisation auxquels vous souhaitez que l'utilisateur ait accès. Pour

obtenir un step-by-step guide, consultez Création d'un nouvel utilisateur dans l'AWS Management Console.

De même, si vous souhaitez créer des utilisateurs administrateurs supplémentaires, vous devez créer un nouvel utilisateur et l'ajouter au groupe d'administrateurs du groupe d'utilisateurs.

Les noms d'utilisateur sont créés en prenant la partie de l'e-mail fourni avant et en ajoutant l'@UUID du cas d'utilisation généré (ou `-admin` dans le cas de l'utilisateur administrateur).

Dans l'onglet Groupes, vous pouvez voir qu'un groupe d'administrateurs et un groupe pour chaque cas d'utilisation ont été automatiquement créés à l'aide du nom du cas d'utilisation (tel que fourni dans l'assistant) et de l'UUID du cas d'utilisation.

Référence des API

Cette section fournit des références d'API pour la solution.

Tableau de bord de déploiement

API REST	Méthode HTTP	Fonctionnalité	Appelants autorisés
`/deployments`	`GET`	Accédez à tous les déploiements.	Jeton JWT authentifié par Amazon Cognito
`/deployments`	`POST`	Crée un nouveau déploiement de cas d'utilisation.	Jeton JWT authentifié par Amazon Cognito
`/deployments/{useCaseId}`	`GET`	Obtient les détails du déploiement pour un seul déploiement.	Jeton JWT authentifié par Amazon Cognito
`/deployments/{useCaseId}`	`PATCH`	Met à jour un déploiement donné.	Jeton JWT authentifié par Amazon Cognito
`/deployments/{useCaseId}`	`DELETE`	Supprime un déploiement donné.	Jeton JWT authentifié par Amazon Cognito
`/model-info/use-case-types`	`GET`	Obtient les types de cas d'utilisation	Jeton JWT authentifié par Amazon Cognito

API REST	Méthode HTTP	Fonctionnalité	Appelants autorisés
		disponibles pour le déploiement	
`/model-info/{useCaseType}/providers`	`GET`	Obtient les fournisseurs de modèles disponibles pour le type de cas d'utilisation donné	Jeton JWT authentifié par Amazon Cognito
`/model-info/{useCaseType}/{providerName}`	`GET`	Obtient IDs les modèles disponibles pour un fournisseur donné et un type de cas d'utilisation	Jeton JWT authentifié par Amazon Cognito
`/model-info/{useCaseType}/{providerName}/{modelId}`	`GET`	Obtient les informations sur le modèle donné, y compris les paramètres par défaut.	Jeton JWT authentifié par Amazon Cognito

> **Note**
>
> Les fichiers OpenAPI et Swagger peuvent également être exportés depuis API Gateway pour faciliter l'intégration à l'API. Reportez-vous à la section Exporter une API REST depuis API Gateway.

Charges utiles POST et PATCH

Vous trouverez ci-dessous un exemple de charge utile POST envoyée au `/deployments` point de terminaison, qui créera un nouveau cas d'utilisation.

```
{
 "UseCaseName": "usecase1",
```

```
"UseCaseDescription": "Description of the use case to be deployed. For display
purposes", // optional
"DefaultUserEmail": "placeholder@example.com", // optional, if not provided, the
Cognito Group and User will not be created
"DeployUI": true, // optional
"VpcParams": {
"VpcEnabled": true,
"CreateNewVpc": false,
// provide these if not creating new vpc
"ExistingVpcId": "vpc-id",
"ExistingPrivateSubnetIds": ["subnet-1", "subnet-2"],
"ExistingSecurityGroupIds": ["sg-1", "sg-2"]
},
"ConversationMemoryParams": {
"ConversationMemoryType": "DynamoDB",
"HumanPrefix": "user", // optional
"AiPrefix": "ai", // optional
"ChatHistoryLength": 10 // optional
},
"KnowledgeBaseParams": {
"KnowledgeBaseType": "Bedrock",
// one of the following based on selected provider
"BedrockKnowledgeBaseParams": {
"BedrockKnowledgeBaseId": "my-bedrock-kb",
"RetrievalFilter": {}, // optional
"OverrideSearchType": "HYBRID" // optional
},
"KendraKnowledgeBaseParams": {
"AttributeFilter": {}, // optional
"RoleBasedAccessControlEnabled": true, // optional
"ExistingKendraIndexId": "12345678-abcd-1234-abcd-1234567890ab",
// provide the following in place of ExistingKendraIndexId if you want the solution to
deploy an index for you
"KendraIndexName": "index",
"QueryCapacityUnits": 1, // optional
"StorageCapacityUnits": 1, // optional
"KendraIndexEdition": "DEVELOPER" // optional
},
"NoDocsFoundResponse": "Sorry, I couldn't find any relevant information for your
query.", // optional
"NumberOfDocs": 3, // optional
"ScoreThreshold": 0.7, // optional
"ReturnSourceDocs": true // optional
},
```

```
 "LlmParams": {
 "ModelProvider": "Bedrock | SAGEMAKER",
 // one of the following based on selected provider
 "BedrockLlmParams": {
 "ModelId": "model-id", // use this for on demand models. Can't use with ModelArn
 "ModelArn": "model-arn", // use this for provisioned/custom models. Can't use with
 ModelId,
 "InferenceProfileId": "profile-id"
 "GuardrailIdentifier": "arn:aws:bedrock:us-east-1:123456789012:guardrail/my-
guardrail", // optional
 "GuardrailVersion": "1" // optional. Required if GuardrailIdentifier provided.
 },
 "SageMakerLlmParams": {
 "EndpointName": "some-endpoint",
 "ModelInputPayloadSchema": {},
 "ModelOutputJSONPath": "$."
 },
 // optional. Passes on arbitrary params to the underlying LLM.
 "ModelParams": {
 "param1": {
 "Value": "value1",
 "Type": "string"
 },
 "param2": {
 "Value": 1,
 "Type": "integer"
 }
 },
 // optional
 "PromptParams": {
 "PromptTemplate": "some template",
 "UserPromptEditingEnabled": true,
 "MaxPromptTemplateLength": 1000,
 "MaxInputTextLength": 1000,
 "DisambiguationPromptTemplate": "some disambiguation template",
 "DisambiguationEnabled": true
 },
 "Temperature": 1.0, // optional
 "Streaming": true, // optional
 "RAGEnabled": true, // optional. Must be true if providing KnowledgeBaseParams above.
 "Verbose": false // optional
 },
 "AgentParams": {
 "AgentType": "Bedrock",
```

```
 "BedrockAgentParams": {
 "AgentId": "agent-id",
 "AgentAliasId": "alias-id",
 "EnableTrace": true
 }
 },
 // optional
 "AuthenticationParams": {
 "AuthenticationProvider": "Cognito",
 "CognitoParams": {
 "ExistingUserPoolId": "user-pool-id",
 "ExistingUserPoolClientId": "client-id" // optional. If not provided, the solution
 will create a client for you in the provided pool
 }
 }

}
```

Pour les mises à jour, la structure est la même que ci-dessus avec quelques réserves :

- Le nom du cas d'utilisation ne peut pas être modifié
- Un cas d'utilisation ne peut modifier les groupes de sécurité et les sous-réseaux qu'une fois qu'il a été déployé dans un VPC. Le VPC lui-même ne peut pas être modifié.
- Si un index Kendra a été créé pour vous en tant que base de connaissances, vous ne pouvez pas modifier la configuration de cet index (par exemple, KendraIndexName) QueryCapacityUnits

Cas d'utilisation partagé APIs

Les points de terminaison de l'API REST suivants sont disponibles pour les cas d'utilisation de Text et Bedrock Agent :

API REST	Méthode HTTP	Fonctionnalité	Appelants autorisés
`/details/{useCaseConfigKey}`	`GET`	Obtient les détails de configuration pour un cas d'utilisation spécifique.	Jeton JWT authentifié par Amazon Cognito

WebSocket API	Fonctionnalité	Appelants autorisés
`/$connect`	Lancez WebSocket la connexion et authentifiez l'utilisateur.	Jeton JWT authentifié par Amazon Cognito
`/$disconnect`	Point de terminaison appelé lorsqu'une WebSocket connexion a été déconnectée.	Jeton JWT authentifié par Amazon Cognito

API des détails des cas d'utilisation

Le point de terminaison de l'API Details récupère des informations sur un cas d'utilisation spécifique :

```
GET /details/{useCaseConfigKey}
```

Ce point de terminaison renvoie les détails de configuration pour un cas d'utilisation spécifique, notamment les paramètres du modèle, les paramètres de la base de connaissances et d'autres informations de déploiement. Il nécessite un jeton JWT authentifié par Amazon Cognito pour l'autorisation.

Cas d'utilisation du texte

WebSocket API	Fonctionnalité	Appelants autorisés
`/sendMessage`	Envoie le message de chat de l'utilisateur au WebSocket pour traitement avec l'expérience LLM configurée.	Jeton JWT authentifié par Amazon Cognito

API REST	Méthode HTTP	Fonctionnalité	Appelants autorisés
`/feedback/{useCaseId}`	`POST`	Soumet les commentaires des utilisateurs pour	Jeton JWT authentifié par Amazon Cognito

API REST	Méthode HTTP	Fonctionnalité	Appelants autorisés
		un cas d'utilisation spécifique.	

Charges utiles d'envoi de messages

Si vous effectuez une intégration directe à l'`/sendMessage`API, vous devez respecter les formats de charge utile de demande et de réponse suivants.

Charge utile de la demande

```
{
 "action": "sendMessage",
 "question": "the message to send to the api",
 "conversationId": "", // If not provided, a new conversation will be created, with the
 conversationId returned in the response. All subsequent messages in that conversation
 (where history is retained), should provide the conversationId there.
 "promptTemplate": "", // Optional. Overrides the configured prompt
 "authToken": "XXXX" // Optional. accessToken from cognito flow. Required for RAG with
 RBAC
}
```

Nom du paramètre	Type	Description
action	`String`	Actuellement, nous prenons uniquement en charge l'action « SendMessage » sur le WebSocket
question	`String`	L'entrée utilisateur à envoyer au LLM
Identifiant de conversation	`String`	Un UUID identifiant la conversation. S'il n'est pas fourni, une nouvelle conversat ion sera créée, avec le ConversationID renvoyé dans la réponse. Tous les

Nom du paramètre	Type	Description
		messages suivants de cette conversation (dans lesquels vous souhaitez qu' history/context ils soient conservés) doivent contenir le ConversationID.
Modèle d'invite	`String`[Facultatif]	Remplace le modèle d'invite pour ce message. S'il est vide ou non fourni, l'invite définie par défaut au moment du déploiement sera utilisée. Les espaces réservés appropriés doivent être spécifiés pour la configuration donnée (c'est-à-dire {history} et {input} pour les déploiements d'IA autres que RAG Sagemaker, avec l'ajout de {context} si vous utilisez RAG pour tous les déploiements).

Nom du paramètre	Type	Description
Jeton d'authentification	`String`[Facultatif]	AccessToken a été obtenu à partir du flux d'authentification cognito. Cela est nécessaire lors de l'appel d'un point de terminaison Websocket de chat configuré pour RAG avec un contrôle d'accès basé sur les rôles (RBAC). La liste de réclamations cognito:groups contenue dans ce jeton JWT est utilisée pour contrôler l'accès aux documents de l'index Kendra. Ce paramètre n'est pas obligatoire pour les cas d'utilisation autres que RAG. Il n'est pas non plus requis pour les cas d'utilisation de RAG pour lesquels le RBAC est désactivé.

Charges utiles de réponse

Réponse à la question

L' WebSocket API répondra avec 1 (si le streaming est désactivé) ou plusieurs (si le streaming est activé) objets JSON structurés comme suit pour chaque requête.

```
{
 "data": "some data",
 "conversationId": "id",
}
```

Nom du paramètre	Type	Description
data	`String`	Une partie de la réponse du LLM si le streaming est activé, ou la totalité de la réponse. Si vous utilisez le streaming , une réponse de ce format dont le contenu des données est END_CONVERSATION sera envoyée pour indiquer la fin de la réponse à une seule question.
Identifiant de conversation	`String`	L'ID de la conversation à laquelle appartient cette réponse SourceDocument.

Réponse au document source

Si vous avez configuré votre cas d'utilisation de RAG pour renvoyer les documents sources, vous recevrez également la charge utile suivante à la fin de chaque réponse pour chaque document source utilisé pour créer la réponse.

```
{
 "sourceDocument": {
 "excerpt": "some excerpt from the",
 "location": "s3://fake-bucket/test.txt",
 "score": 0.500,
 "document_title": null,
 "document_id": null,
 "additional_attributes": null
 },
 "conversationId": "some-id"
}
```

Nom du paramètre	Type	Description
extrait	`String`	Extrait du document source.

Nom du paramètre	Type	Description
location	`String`	Emplacement du document source. Cela dépendra des sources de données utilisées et du type de base de connaissances, mais il peut s'agir de choses comme s3 URIs ou de sites Web.
score	`Number` | `String`	La certitude que le document correspond à la question posée. Il s'agira d'un float de 0 à 1 pour Bedrock et d'une chaîne (par exemple HIGH, LOW, etc.) pour Kendra.
titre_document	`String`	Titre du document source renvoyé. Disponible uniquement lorsque vous utilisez Kendra.
identifiant_document	`String`	ID du document source renvoyé. Disponible uniquement lorsque vous utilisez Kendra.
attributs_supplémentaires	`String`	Ce champ contiendra tous les attributs supplémentaires du document tels que personnal isés dans votre base de connaissances lors de l'ingesti on.
Identifiant de conversation	`String`	L'ID de la conversation à laquelle appartient cette réponse SourceDocument.

Charge utile de l'API Feedback

Vous trouverez ci-dessous un exemple de charge utile POST envoyée au `/feedback/{useCaseId}` point de terminaison, qui soumettra les commentaires des utilisateurs pour un cas d'utilisation spécifique :

```
{
  "useCaseRecordKey": "12345678-12345678",
  "conversationId": "12345678-1234-1234-1234-123456789012",
  "messageId": "12345678-1234-1234-1234-123456789012",
  "feedback": "positive",
  "feedbackReason": ["accurate", "helpful"],
  "comment": "This response was very helpful.",
  "rephrasedQuery": "What are the key features of Amazon Bedrock?",
  "sourceDocuments": [
    "s3://bucket-name/document1.pdf",
    "s3://bucket-name/document2.pdf"
  ]
}
```

Cas d'utilisation de Bedrock Agent

WebSocket API	Fonctionnalité	Appelants autorisés
`/invokeAgent`	Envoie le message de l'utilisateur au WebSocket pour traitement avec l'agent configuré.	Jeton JWT authentifié par Amazon Cognito

Charges utiles d'InvokeAgent

Si vous effectuez une intégration directe avec le`/invokeAgent API`, vous devez respecter les formats de charge utile de demande et de réponse suivants.

Charge utile de la requête

```
{
"action": "invokeAgent",
"inputText": "User query to the agent",
```

```
"conversationId": "", // Optional. Empty conversationId implies a new conversation.
 When not provided, a new conversationId will be created and returned with the
 response. All subsequent messages in the same conversation should provide the same
 conversationId (i.e. chat memory/history is maintained).
"authToken": "XXXX" // Optional. accessToken from cognito flow. If provided, it needs
 to be a valid JWT token associated with the user
}
```

Nom du paramètre	Type	Description
action	`String`	Nous soutenons uniquement l'`invokeAgent` action sur le WebSocket.
Texte de saisie	`String`	Entrée utilisateur à envoyer au LLM.
Identifiant de conversation	`String[Optional]`	Un UUID qui identifie la conversation de manière unique. Si vous ne fournisse z pas cette valeur, la solution crée une nouvelle conversat ion et renvoie le Conversat ionId dans la réponse. Tous les messages suivants de cette conversation (dans lesquels vous souhaitez conserver l'historique et le contexte) contiennent le ConversationID.
Jeton d'authentification	`String[Optional]`	AccessToken a été obtenu à partir du flux d'authent ification Amazon Cognito. Ce paramètre n'est pas obligatoi re. Si vous le fournissez, le jeton JWT sera validé. Cela

Nom du paramètre	Type	Description
		permet de faciliter l'extension de cette solution.

Charges utiles de réponse

Réponse à la question

L' WebSocket API répondra avec un (si le streaming est désactivé) ou plusieurs (si le streaming est activé) objets JSON structurés comme suit pour chaque requête.

```
{
 "data" "some data",
 "conversationId": "id",
 }
```

Nom du paramètre	Type	Description
data	`String`	La réponse à l'invocation de l'agent.
Identifiant de conversation	`String`	L'identifiant de la conversation.

Référence

Cette section inclut des informations sur la collecte de données pour cette solution, des pointeurs vers des ressources connexes et une liste des créateurs qui ont contribué à cette solution.

Fournisseurs de LLM pris en charge

La solution peut s'intégrer aux fournisseurs de LLM suivants :

1. Amazon Bedrock
   - Documents : https://aws.amazon.com/bedrock/
   - Modèles pris en charge :
     - Amazon
       - Nova Lite
       - Nova Micro
       - Nova Pro
     - AI21 Laboratoires
       - Jamba 1.5 Mini
       - Jamba 1.5 Large
     - Anthropic
       - Claude v3 Haïku
       - Claude v3.5 Sonnet
       - Claude v3.7 Sonnet (grâce à l'utilisation de profils d'inférence)
     - Cohere
       - Command R
       - Command R+
     - Deepseek
       - Deepseek-R1 (grâce à l'utilisation de profils d'inférence)
     - Meta
       - Llama 3
       - Llama 3.2 (grâce à l'utilisation de profils d'inférence)
     - Mistral AI

        - Mistral 7B Instruct
        - Mistral 8x7B Instruct
    - Inférence entre régions
        - Possibilité d'utiliser des profils d'inférence définis dans la même région que le tableau de bord de déploiement
2. Amazon SageMaker AI
    - Documents : https://aws.amazon.com/sagemaker/
    - Modèles pris en charge : modèles de texte en texte

Pour connaître les derniers paramètres du modèle, les meilleures pratiques et les utilisations recommandées, reportez-vous à la documentation des fournisseurs de modèles.

Collecte des données

Cette solution envoie des métriques opérationnelles à AWS (les « données ») concernant l'utilisation de cette solution. Nous utilisons ces données pour mieux comprendre comment les clients utilisent cette solution et les services et produits associés. La collecte de ces données par AWS est soumise à l'avis de confidentialité d'AWS.

Collaborateurs

- Tarek Abdounabi
- Majid Arbash
- George Bearden
- Mukit Bin Momin
- Michael Connor
- Johnny Duval
- Nihit Kasabwala
- Ahern Knox
- Simon Krol
- Michael Lin
- Tim Mekari

- Ibrahim Mohamed
- Omar Radwan Mohsen
- James Nixon
- Dekshitha Ravikumar
- Jae Shim
- Ajay Swamy
- Mohammed Taha
- Reet Takkar
- Dimitri Tchikatilov
- Couronne Jason
- Kamyar Ziabari

Révisions

Date de publication : octobre 2023 (dernière mise à jour : janvier 2025)

Consultez le fichier ChangeLog.md dans le GitHub référentiel pour voir toutes les modifications et mises à jour notables apportées au logiciel. Le journal des modifications fournit un enregistrement clair des améliorations et des correctifs pour chaque version.

Notifications

Il incombe aux clients de procéder à une évaluation indépendante des informations contenues dans le présent document. Ce document : (a) est fourni à titre informatif uniquement, (b) représente les offres de produits et les pratiques actuelles d'AWS, qui sont susceptibles d'être modifiées sans préavis, et (c) ne crée aucun engagement ni aucune garantie de la part d'AWS et de ses filiales, fournisseurs ou concédants de licence. Les produits ou services AWS sont fournis « tels quels » sans garanties, déclarations ou conditions d'aucune sorte, qu'elles soient explicites ou implicites. Les responsabilités et obligations d'AWS à l'égard de ses clients sont régies par les accords AWS, et ce document ne fait partie d'aucun accord conclu entre AWS et ses clients et ne les modifie pas.

Generative AI Application Builder sur AWS est concédé sous licence selon les termes de la licence Apache version 2.0.

Important

Generative AI Application Builder sur AWS vous permet de créer et de déployer des applications d'intelligence artificielle générative sur AWS en utilisant le modèle d'IA générative de votre choix, y compris des modèles d'IA générative tiers que vous pouvez choisir d'utiliser et qu'AWS ne possède pas ou sur lesquels AWS n'a aucun contrôle (« modèles d'IA générative tiers »).

Votre utilisation des modèles d'IA générative tiers est régie par les conditions qui vous ont été fournies par les fournisseurs tiers de modèles d'IA générative lorsque vous avez acquis votre licence d'utilisation (par exemple, leurs conditions d'utilisation, leur contrat de licence, leur politique d'utilisation acceptable et leur politique de confidentialité).

Il vous incombe de veiller à ce que votre utilisation des modèles d'IA générative tiers soit conforme aux conditions qui les régissent, ainsi qu'à toutes les lois, règles, réglementations, politiques ou normes qui s'appliquent à vous.

Vous êtes également responsable de faire votre propre évaluation indépendante des modèles d'IA générative tiers que vous utilisez, y compris leurs résultats et de la manière dont les fournisseurs de modèles d'IA générative tiers utilisent les données qui pourraient leur être transmises en fonction de votre déploiement. AWS ne fait aucune déclaration ni ne donne aucune garantie concernant les modèles d'IA générative tiers, qui constituent du « contenu tiers » au sens de votre accord avec AWS. Generative AI Application Builder sur AWS vous est proposé en tant que « contenu AWS » dans le cadre de votre contrat avec AWS.

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.