Oracle Generative AI Agent

Service d'IA Générative pour OCI

Ce cours présente les concepts fondamentaux d'Oracle Generative AI Agents, un service entièrement géré qui combine la puissance des grands modèles de langage avec un système de récupération intelligent pour créer des réponses contextuellement pertinentes.

Table des matières

1 Introduction aux Oracle Generative AI Agents

Oracle Generative AI Agents est un service entièrement géré qui combine la puissance des grands modèles de langage (LLM) avec un système de récupération intelligent visant à créer des réponses contextuellement pertinentes en effectuant des recherches dans votre base de connaissances.

Les agents d'IA générative représentent une évolution significative dans l'utilisation des grands modèles de langage. Contrairement aux implémentations basiques de LLM, ces agents intègrent des capacités de raisonnement, de planification et d'action qui leur permettent d'exécuter des tâches complexes de manière autonome.

Exemple

Supposons que nous demandions à cet agent IA : « Réservez-moi un vol pour Las Vegas ainsi qu'une chambre au Hilton Hotel. »
L'agent va :

- Comprendre et interpréter la requête
- Déterminer les prochaines étapes à entreprendre
- Récupérer les données des magasins de données
- Donner une réponse ou exécuter une action

La réponse pourrait être : « Votre voyage est réservé », très probablement après avoir effectué les actions nécessaires.

Point important

Les agents sont des applications de grands modèles de langage, packagées et validées, prêtes à être utilisées immédiatement. OCI Generative AI Agents prend en charge plusieurs façons d'intégrer vos données, où vous et vos clients peuvent interagir avec vos données en utilisant une interface de chat ou une API.

2 Architecture Globale

L'architecture d'Oracle Generative AI Agent suit un modèle sophistiqué qui intègre plusieurs composants essentiels pour délivrer des réponses intelligentes et contextuelles.

2.1 Composants Principaux

2.1.1 Interface Utilisateur

Le parcours commence par l'interface, qui est le point où l'utilisateur interagit avec l'agent IA. Cette interface peut prendre plusieurs formes :

- Chatbot intégré
- Application web personnalisée
- Interface vocale
- Toute application où l'utilisateur saisit une requête ou une commande

2.1.2 Entrées du Modèle de Langage

Le système fournit diverses entrées au grand modèle de langage :

Mémoire à court/long terme : Elle peut fournir le contexte des interactions passées, permettant la continuité et la pertinence dans les conversations.

Outils : Vous pouvez intégrer différents outils externes, par exemple, différentes API, bases de données, ou systèmes tiers pour améliorer les capacités du modèle.

Prompt : Il contient la requête ou la tâche spécifique fournie par l'utilisateur guidant l'IA sur la façon de générer des réponses.

2.1.3 Le Grand Modèle de Langage

Au cœur du système se trouve le grand modèle de langage qui effectue quatre opérations clés :

Opération	Description	Fonction
Raisonnement	Analyse logique	Analyse les entrées pour générer des réponses logiques et cohérentes
Action	Détermination d'actions	Détermine les actions basées sur la tâche, par exemple, interroger des bases de données ou appeler différentes API
Persona	Maintien de la cohérence	Maintient un ton, un style et un comportement cohérents alignés avec la marque ou le cas d'usage
Planification	Organisation stratégique	Organise stratégiquement les réponses ou actions, en particulier dans les flux de travail multi-étapes

2.2 Accès aux Bases de Connaissances

Le LLM peut également accéder aux bases de connaissances externes, telles que les bases de données ou les référentiels de documents, pour enrichir ses réponses avec des informations précises et à jour. Cette capacité s'appuie sur la technique RAG (Retrieval-Augmented Generation) que vous avez déjà vue dans les leçons précédentes.

Point important

Cette architecture permet à l'agent de dépasser ses données d'entraînement internes et d'accéder à des informations actualisées et spécifiques au domaine.

2.3 Boucle de Rétroaction

Basé sur toutes les entrées traitées, le raisonnement et les connaissances intégrées, le LLM génère une réponse. Cette réponse est adaptée à la requête et au contexte fournis par l'utilisateur.

La sortie générée par l'agent IA peut être réinjectée dans sa mémoire à court terme, permettant d'améliorer les réponses dans les interactions en cours. Ainsi, il existe globalement une boucle de rétroaction qui assure l'apprentissage continu et l'amélioration des performances.

Cette architecture garantit que l'Oracle Generative AI Agent délivre des réponses hautement intelligentes, contextuelles et exploitables en exploitant les entrées utilisateur, les outils externes et les capacités de raisonnement robustes. Elle est conçue pour la scalabilité, l'adaptabilité et l'efficacité dans les applications d'entreprise.

3 Concepts Fondamentaux

Nous allons maintenant explorer les concepts de base qui permettent aux agents de fournir des interactions intelligentes et contextuellement pertinentes.

3.1 Modèle d'IA Générative

Au cœur d'OCI Generative Agents se trouve le modèle d'IA générative. Il s'agit d'un grand modèle de langage (LLM) entraîné sur de vastes ensembles de données pour générer du texte similaire à celui d'un humain.

Ce modèle traite de nouvelles entrées pour produire des réponses cohérentes et contextuellement appropriées, permettant la compréhension et la génération de langage naturel.

3.2 L'Agent

L'agent lui-même est un système autonome construit sur le LLM. Il comprend et génère du texte tout en facilitant les interactions en langage naturel.

Point important

OCI prend en charge les agents RAG qui se connectent aux sources de données, récupèrent les informations pertinentes et améliorent les réponses du modèle avec ces données, ce qui garantit des sorties plus précises et pertinentes.

3.2.1 Critères de Performance RAG

Lors de l'utilisation d'agents RAG, les modèles doivent performer avec une haute capacité de réponse et un bon ancrage :

Capacité de réponse (Answerability) : Le modèle peut générer des réponses pertinentes à différentes requêtes utilisateur.

Ancrage (Groundedness) : Les réponses générées par le modèle doivent être traçables à différentes sources de données, garantissant leur fiabilité et leur exactitude.

3.3 Hiérarchie des Données

Pour fonctionner efficacement, un agent accède aux données à travers une hiérarchie structurée comprenant trois niveaux :

Niveau	Composant	Description

Niveau 1	Data Store	Le référentiel où résident les données, comme les buckets de stockage d'objets ou les bases de données
Niveau 2	Data Source	Fournit les détails de connexion au data store, permettant à l'agent d'accéder et de récupérer les données
Niveau 3	Knowledge Base	Le système de stockage vectoriel qui ingère les données de la source de données, les organisant pour une récupération et utilisation efficaces par l'agent

Cette structure garantit que les agents peuvent accéder de manière transparente et utiliser les informations nécessaires pour générer des réponses informées.

4 Options de Sources de Données

Oracle fournit plusieurs options de données pour rendre vos informations accessibles aux agents d'IA générative.

4.1 Données de Stockage d'Objets

Vous pouvez télécharger directement des fichiers de données vers OCI Object Storage, permettant au service d'ingérer automatiquement les données. Il s'agit d'une option gérée par le service, ce qui signifie que le service prend en charge cette partie d'ingestion.

4.2 Données OpenSearch

Vous pouvez apporter vos propres données ingérées et indexées depuis OCI Search avec OpenSearch pour que les agents les utilisent.

4.3 Magasin de Vecteurs Oracle Database

Vous pouvez apporter vos propres embeddings vectoriels depuis une base de données Oracle 23ai ou une base de données autonome 23ai vers les agents d'IA générative.

Point important

Dans ce cours, nous discuterons des options de stockage d'objets et Oracle Database 23ai.

4.4 Processus d'Ingestion des Données

L'ingestion des données est le processus d'extraction des données des documents sources, de leur transformation en un format structuré approprié pour l'analyse et de leur stockage dans la base de connaissances.

Cette étape est cruciale pour préparer les données brutes afin que l'agent puisse y accéder et les utiliser efficacement lors des interactions.

5 Concepts Avancés

5.1 Session

Une session représente la conversation interactive initiée par un utilisateur, maintenant le contexte tout au long de l'échange pour assurer des réponses cohérentes et pertinentes.

5.2 Point de Terminaison d'Agent (Agent Endpoint)

Il s'agit d'un point d'accès spécifique qui permet à l'agent de communiquer avec des systèmes ou services externes. Il facilite également l'échange de données, permettant à l'agent de récupérer ou d'envoyer des informations selon les besoins pour exécuter ses fonctions efficacement.

5.3 Traçabilité (Trace)

Cette fonctionnalité suit et affiche l'historique d'une conversation de chat, incluant à la fois les prompts utilisateur et les réponses de l'agent. Cette fonctionnalité est précieuse pour :

- Surveiller les interactions
- Comprendre le processus de prise de décision
- Assurer la transparence dans les opérations de l'agent

5.4 Citation

La citation fait référence à la source d'information utilisée dans la réponse de l'agent. L'agent RAG fournit des citations pour chaque réponse, incluant des détails comme le titre, le chemin externe, l'ID du document et les numéros de page.

Cela garantit que les utilisateurs peuvent tracer les réponses jusqu'à leur source originale, améliorant la confiance et la responsabilité.

5.5 Modération de Contenu

Point important

Il s'agit d'une fonctionnalité conçue pour détecter et filtrer le contenu nuisible dans les prompts utilisateur et les réponses générées. Elle se concentre sur l'identification et l'atténuation de divers types de préjudices, notamment :

- Haine et harcèlement
- Préjudice auto-infligé
- Préjudice idéologique
- Exploitation

La modération peut être appliquée uniquement au prompt utilisateur, uniquement à la réponse générée, ou aux deux, garantissant que les interactions restent sûres et respectueuses.

6 Directives pour le Stockage d'Objets

Nous savons déjà que nous téléchargeons des fichiers de données vers OCI Object Storage et laissons les agents d'IA générative ingérer automatiquement les données. Voyons maintenant quelques directives sur l'utilisation du stockage d'objets comme source de données.

6.1 Configuration de Base

Paramètre	Spécification
Association bucket	Chaque source de données est associée à un seul bucket (un seul bucket autorisé par source de données)
Formats supportés	PDF et fichiers texte uniquement
Taille maximale	Chaque fichier ne doit pas dépasser 100 MB
Éléments graphiques	Les fichiers PDF peuvent inclure des images, graphiques et tableaux de référence, mais ces éléments ne doivent pas dépasser 8 MB

6.2 Spécifications pour les Graphiques

Point important

Vous devez vous assurer que les graphiques sont bidimensionnels avec des axes étiquetés. Le modèle peut interpréter et répondre aux questions sur ces graphiques sans aucune préparation supplémentaire.

6.3 Tableaux de Référence

Vous pouvez également utiliser des tableaux de référence avec plusieurs lignes et colonnes. L'agent peut lire et interpréter efficacement de tels tableaux.

6.4 Gestion des Liens Hypertexte

Tous les liens hypertexte présents dans les documents PDF sont extraits et affichés comme des liens cliquables dans les réponses de chat.

6.5 Préparation Flexible

Si vos données ne sont pas encore prêtes, vous pouvez toujours créer un dossier vide pour la source de données et le peupler plus tard. Cette approche permet essentiellement de configurer la source de données à l'avance et d'ingérer les données une fois qu'elles sont disponibles.

En suivant ces directives de stockage d'objets, vous pouvez vous assurer que vos données sont correctement préparées et accessibles pour les agents d'IA générative d'Oracle, conduisant à des interactions pilotées par l'IA plus efficaces et efficientes.

7 Directives pour Oracle Database

Passons maintenant aux directives Oracle Database concernant les agents d'IA générative.

7.1 Configuration Préalable

Point important

Les agents d'IA générative ne gèrent pas la base de données, vous devez donc configurer votre base de données existante pour que les agents d'IA générative puissent s'y connecter.

7.2 Structure de Table Requise

Vous devez créer une table Oracle Database 23ai avec les champs suivants :

Champ	Type	Description
DOCID	Obligatoire	Identifiant unique du document
body	Obligatoire	Chunks (fragments) des données textuelles
vector	Obligatoire	Vecteur généré à partir du body en utilisant un modèle d'embedding
CHUNKID	Optionnel	Identifiant du chunk
URL	Optionnel	URL source du document
title	Optionnel	Titre du document
page numbers	Optionnel	Numéros de pages

7.3 Fonction de Recherche Vectorielle

Vous devrez configurer une fonction de base de données qui peut retourner les résultats de recherche vectorielle pour chaque requête.

Une fonction est un sous-programme qui peut prendre des paramètres et retourner une valeur.

7.3.1 Paramètres Requis

Pour les entrées requises, vous avez des paramètres tels que :

- `p_query` : La requête de recherche
- `top_k` : Le nombre de résultats à retourner

Exemple

Nom de fonction : `retrieval_func_ai`
Paramètres d'entrée :

- `p_query` : Requête utilisateur
- `top_k` : Nombre de résultats souhaités

7.4 Exigences d'Alignement des Modèles

Point important

Vous devez vous assurer que le modèle d'embedding utilisé pour le champ de requête de la fonction correspond au modèle d'embedding qui a transformé le contenu du corps de la table de base de données en embeddings vectoriels.

Cela signifie que le modèle d'embedding utilisé dans la requête (par exemple, Cohere embed multilingual v3) doit correspondre au modèle d'embedding utilisé pour générer la colonne text_vec dans la base de données.

7.5 Structure de Retour

Le champ de retour de la fonction doit s'aligner avec les champs requis de la table :

- DOCID
- body
- score (calculé)

Si les noms des champs de retour de la fonction diffèrent des noms des champs de la table, vous pouvez utiliser des alias dans la fonction.

7.6 Fonctionnement de la Recherche Vectorielle

Le processus de recherche suit ces étapes :

1. La requête accède à une table qui contient des embeddings vectoriels
2. La table inclut des colonnes telles que DOCID, body, et text_vector
3. Le système calcule la distance vectorielle en utilisant la similarité cosinus ou la distance euclidienne
4. Cette distance vectorielle représente la distance entre le vecteur de requête et les embeddings de documents
5. La requête récupère les top_k lignes triées par score de similarité en ordre décroissant
6. Cela garantit que les résultats les plus pertinents sont retournés à l'utilisateur

La fonction retourne un SYS_REFCURSOR avec les champs DOCID, body, et score, permettant un accès efficace aux données les plus pertinentes pour la requête utilisateur.

8 Flux de Travail de Création d'Agent

Maintenant que vous êtes familiarisé avec les directives, examinons rapidement les différents concepts vus précédemment et essayons de comprendre le flux de travail global de création d'un agent.

8.1 Étape 1 : Création de la Base de Connaissances

8.1.1 Base de Connaissances avec Stockage d'Objets

Vous commencez par la base de connaissances utilisant le stockage d'objets comme data store. Vous pouvez voir ici différents objets dans les buckets définis comme source de données pour cette base de connaissances.

Le processus comprend :

1. Fournir le nom, compartiment et autres informations nécessaires
2. Choisir le type de data store
3. Activer la recherche hybride (combinaison de recherche lexicale et sémantique)
4. Spécifier votre source de données stockée dans le bucket
5. Créer la base de connaissances
6. Choisir de démarrer le job d'ingestion immédiatement ou de le créer manuellement plus tard

8.1.2 Base de Connaissances avec Oracle 23ai

Pour Oracle 23ai, le processus est légèrement différent :

1. Fournir nom, compartiment et autres informations nécessaires
2. Dans le type de data store, choisir Oracle AI vector search
3. Configurer la connexion à l'outil de base de données
4. Lister cette connexion et fournir la fonction de recherche vectorielle
5. Cliquer sur Créer

8.2 Étape 2 : Création de l'Agent

Après avoir créé la base de connaissances, l'étape suivante consiste à créer un agent :

1. Fournir les informations nécessaires
2. Définir le message de bienvenue et les instructions pour la génération RAG si nécessaire
3. Choisir l'une des bases de connaissances créées précédemment (Oracle 23ai ou stockage d'objets)
4. Finaliser la création de l'agent

8.3 Étape 3 : Création du Point de Terminaison

Une fois l'agent créé, vous pouvez créer un endpoint :

> Un endpoint est un point d'accès spécifique qui permet à l'agent de communiquer avec des systèmes externes ou différents services.

Lors de la création, vous pouvez configurer :

— Modération de session

— Traçabilité

— Citations

8.4 Étape 4 : Interaction avec l'Agent

La dernière étape consiste à chatter avec l'agent en utilisant l'endpoint d'agent que vous venez de créer. Dans l'interface, vous pouvez voir :

Citations : La traçabilité ou l'endroit d'où vos informations sont récupérées (groundedness)

Traces : Le maintien de la requête ainsi que des réponses pour assurer la transparence

Point important

Nous verrons toutes ces étapes individuellement dans les leçons à venir avec deux démonstrations : une sur la base de connaissances de stockage d'objets et l'autre sur la base de connaissances Oracle Database 23ai.

9 Limites de Ressources par Défaut

Il est important de connaître les limites de ressources par défaut du service :

Ressource	Limite par Défaut
Agents par compartiment	5 agents maximum
Bases de connaissances par compartiment	10 bases de connaissances maximum
Sources de données par base de connaissances	3 sources maximum
Endpoints par agent	2 endpoints maximum
Sessions actives par endpoint	100 sessions maximum
Taille de fichier individuel	100 MB maximum
Éléments graphiques PDF	8 MB maximum

Point important

Gardez à l'esprit que vous pouvez toujours faire une demande pour augmenter ces limites selon vos besoins spécifiques.

10 Conclusion

Ce cours a présenté un aperçu complet d'Oracle Generative AI Agents et de son architecture. Nous avons exploré :

— L'architecture globale et les composants principaux du service

— Les concepts fondamentaux des agents d'IA générative
— Les différentes options de sources de données (stockage d'objets et Oracle Database)
— Les directives spécifiques pour chaque type de source de données
— Le flux de travail complet de création d'un agent
— Les limites de ressources à considérer

Oracle Generative AI Agents représente une solution puissante pour créer des applications d'IA conversationnelle robustes, capables de fournir des réponses contextuelles et fiables en s'appuyant sur vos propres données d'entreprise.

Les agents d'IA générative sont des applications LLM packagées, validées et prêtes à l'emploi, conçues pour la scalabilité, l'adaptabilité et l'efficacité dans les environnements d'entreprise.

Dans les prochaines leçons, nous approfondirons avec des démonstrations pratiques montrant comment implémenter concrètement ces concepts avec les deux types de bases de connaissances présentés.