DIY : Déployer **ChatGPT** sur une base documentaire privée en toute sécurité-confidentialité

Objectif : interroger une base documentaire volumineuse en langage naturel et obtenir une réponse synthétique ainsi qu'une liste de documents de référence, le tout avec un excellent niveau de fiabilité.

La recherche au sein d'une base documentaire est toujours une entreprise chronophage, aléatoire et, en général, frustrante. La plupart du temps, les moteurs de recherche (en ligne ou privés) génèrent, à partir de mots clés, de longues listes de documents au sein desquelles l'utilisateur doit encore identifier les passages qui lui seront, éventuellement, utiles. Deux récentes innovations ont permis de grandes avancées dans ce domaine. D'un côté les moteurs de recherches nouvelles génération et les bases de données vectorielles permettent la recherche par analyse du sens plutôt que par mots clefs. Et de l'autre ChatGPT et ses équivalents ont apporté une première couche de compréhension de questions et de restitution de réponses en langage naturel avec des performances jusqu'à présent hors de portée. Ensemble ces deux technologies fournissent une solution incomparablement plus performante: ces moteurs exécutent une recherche plus complexe, car formulée en langage naturel, de façon plus efficace, car ils sont en mesure de formuler une synthèse fiable et de citer leurs sources..

Ces capacités sont utiles à tous les professionnels qui utilisent une base de documents volumineuse. Les professions du droit sont au premier plan des utilisateurs potentiels, et de nombreux articles ont déjà décrits dans le détail comment ChatGPT les aide dans leurs travaux. Cependant, tant leur coût, modéré, qu'un temps de développement, réduit, mettent déjà ces systèmes à disposition d'un large panel de situations et de professions. Et il est vraisemblable que nombre de ces dernières sauront également bénéficier des apports de ces outils. Pour les professions du risque, on peut imaginer des bases de données de rapports d'audit, de papiers de travail et de tests, de lois et de règlements, de procédures, de normes, de documentations etc. Ces bases de données, rendues accessibles au travers de ChatGPT (ou équivalent), de façon sécurisée et confidentielle, permettraient aux contrôleurs et auditeurs internes de rechercher des types de test, des exemples de recommandations, des procédures… le tout en langage naturel, sur des années d'archives et en un instant. Voilà bon nombre de problèmes de gestion documentaire et de capitalisation des connaissances résolus « relativement » simplement.

Du reste, certains grands groupes expérimentent déjà ces outils. C'est la raison pour laquelle l'IFACI a développé un démonstrateur permettant d'interroger le corpus documentaire de l'IIA (IPPF, Position Paper et Practice Guide): nous souhaitons explorer avec vous les capacités de ces outils ainsi que les conditions de leur mise en œuvre. Ce ChatBot est disponible ici. Il peut être utilisé de façon sécurisée et anonyme par tous les adhérents de l'IFACI. S'il donne satisfaction, il sera progressivement enrichi de contenus plus opérationnels. C'est sur la base de ce démonstrateur que nous mettons à votre disposition les informations suivantes.

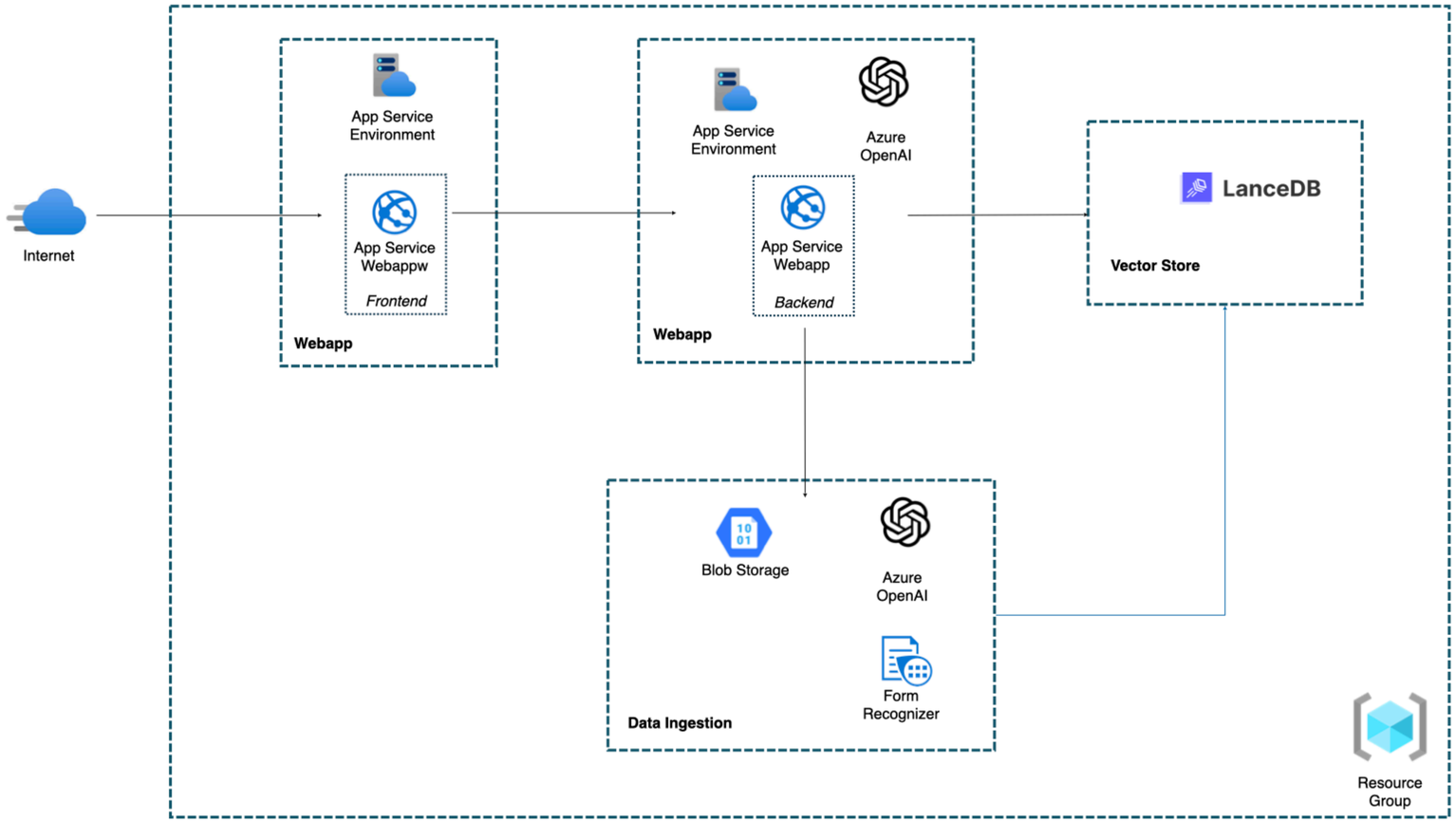
ifaci

A/ Les principes et l'architecture de la solution

Le démonstrateur est construit sur une instance fermée Microsoft Azure (gérée par l'IFACI) et utilise le moteur ChatGPT 4 d'OpenAI.

Les points clés :

- L'accès à ChatGPT 4 peut être sécurisé via un système d'authentification SSO standard (en l'occurence, celui de l'IFACI).
- L'accès à ChatGPT 4 est anonyme grâce à l'architecture de service Microsoft Azure.
- Les documents sont stockés sur un espace sécurisé Microsoft Azure.
- Les requêtes ("prompt") ne sont pas intégrées dans la version publique de ChatGPT.



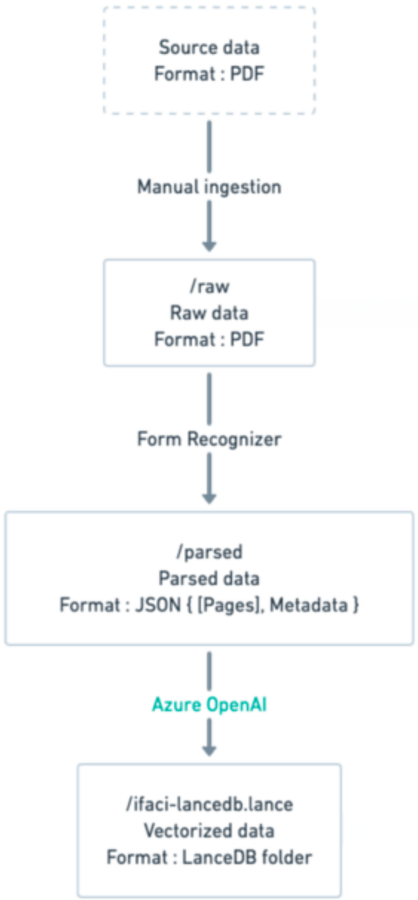
ifaci

B/ Les documents et les traitements appliqués

Afin de créer une instance spécifique de l'outil ChatGPT, il faut que les données sources spécifiques, préalablement inventoriées, soient intégrées dans une base de données créée pour l'occasion. On parle d'ingestion des données.

Nous avons intégré l'ensemble des documents PDF de l'IPPF (International Professional Practice Framework de l'IIA) au sein du démonstrateur GAIA. Le processus d'ingestion nécessite de découper les documents en fragments de texte utilisables par la base de données. On parle de « chunks » (morceaux) qui seront stockés dans la base de données « vectorielle » (car cette dernière stocke les informations sous forme de vecteurs à n dimensions).

Des développements spécifiques peuvent être nécessaires pour ajouter des fonctionnalités au moteur ChatGPT. Dans le cas de Gaïa, nous avons choisi d'ajouter la citation des sources aux réponses générées par le système.



Les points clés :

- Les documents sont découpés en de courts fragments "chunks" porteurs de sens.
- Les parties des documents « vide de sens » sont supprimées, ainsi que les répétitions qui pourraient perturber l'équilibre (le poids) des informations stockées dans la base vectorielle.
- Après vectorisation des « chunks », les données sont stockées pour traitement par ChatGPT dans une base de données vectorielle spécifique et privée.
- Ajout des sources et des titres des documents aux réponses générées par défaut.

C/ Les réglages du système

Qualité des réponses

Après ingestion des données le système doit être testé sur des questions dont les réponses sont connues. Il sera alors possible d'évaluer la pertinence et la qualité des réponses. Le réglage des paramètres de ChatGPT pourra alors être affiné.

Les points clés :

- Le système interprète et « dérive » depuis les orignaux contenus dans la base de données. Des paramètres permettent de contrôler la dérive du système tel que : la température ou le nombre d'éléments maximum utilisés pour formuler une réponse.
- Quels que soient les paramètres, une référence vers le document original sera nécessaire si l'on souhaite préserver l'intégrité de la réponse car la réponse formulée par le système diffère de la source. Dans le cas d'une recherche sur des normes, par exemple, où la précision du vocabulaire est essentielle, la consultation des sources citées est requise.

Besoin en termes de charge

Afin d'avoir un système fonctionnel, il est nécessaire d'estimer la charge maximale auquel celui-ci sera confrontée. En effet la charge que peut supporter le système (en termes de nombre de requête par seconde, par exemple) est un paramètre de l'environnement Microsoft Azure qui ne peut pas être adapté instantanément. Il est par ailleurs, à ce stade, contingenté par Microsoft.

D/ Les coûts et les délais

L'IPPF comporte environ 200 documents. Il s'agit d'une base de données peu volumineuse et simple à traiter car intégralement au format PDF (par opposition, des vidéos, des images ou des fichiers audios seraient plus complexes à traiter). L'analyse du système, l'ingestion des documents, le développement du frontend et le réglage du système ont représenté 5 semaines de délai et un investissement d'environ 50 k€. En interne, la charge de suivi et de test n'a pas excédée 5 j/h. Cela étant, Gaïa reste un simple démonstrateur et la qualité des réponses n'a pas encore été correctement testée à ce stade.

E/ Q/R

- **Y a-t-il une dérive des réponses du système dans le temps ?**
  - La dérive peut être controlée par différent facteurs.
    - La température permet de contrôler le niveau de « créativité » des réponses générées par le modèle.
    - Une surveillance du modèle est nécessaire (prise en compte des retours utilisateurs pour mise à jour).
    - L'excès de personnalisation d'une instance d'IA peut provoquer des dérives difficiles à contrôler du fait de l'éloignement du modèle d'origine.

- **Comment faire évoluer la base documentaire ?**
  - Pour chaque évolution de la base documentaire, il est nécessaire de prévoir:
    1. l'ingestion des nouvelles données : leur préparation et leur intégration dans la base de données vectorielle.
    2. l'évaluation du modèle : après la mise à jour, la performance doit être re-testée du modèle avec des données de test pour s'assurer qu'il fonctionne toujours de manière satisfaisante.

- **Quelles natures de contenus peut-on imaginer insérer dans la base documentaire ?**
  - Il est possible d'intégrer différents types de contenus, comme du son, des images ou des vidéos qui, interprétés, pourront alimenter la base de données vectorielle, même si les modèles actuels sont plus performants sur du texte.

- **Quelles utilisations de ce type de systèmes peut-on imaginer pour l'audit interne ?**
  - Un service d'audit interne pourrait, par exemple, utiliser un système incluant des bases distinctes et complémentaires :

    1. Une base de données de type Gaïa, contenant les normes, les bonnes pratiques, la démarche qualité, des exemples de livrables etc.
    2. Une base de données contenant les données de l'audit interne de l'organisation (rapport, feuille de travail, compte rendu d'entretien…)
    3. Une base de données contenant toutes les autres données disponibles de l'organisation (procédures, architecture SI, bilan …)
    4. Le data lake de l'organisation (pour formuler des requêtes générées par ChatGPT L'utilisation conjointe de ces bases de données pourrait permettre un accompagnement des auditeurs et du département d'audit tout au long des processus opérationnels.

- **Peut-on traiter des données avec ce type d'outils ?**
  - ChatGPT-4 et leurs concurrents sont des modèles de langage, principalement conçus pour interpréter et générer du texte de manière contextuelle. Ces modèles ne sont pas spécifiquement conçus pour effectuer des tâches de traitement de données ou d'analyse de données. Cependant, des modules ad hoc permettent aujourd'hui à ChatGPT, par exemple, de proposer des services d'analyse de données. Ces outils (et notamment les questions de sécurité des requêtes et des données) n'ont pas été traités dans le cadre du démonstrateur Gaïa.