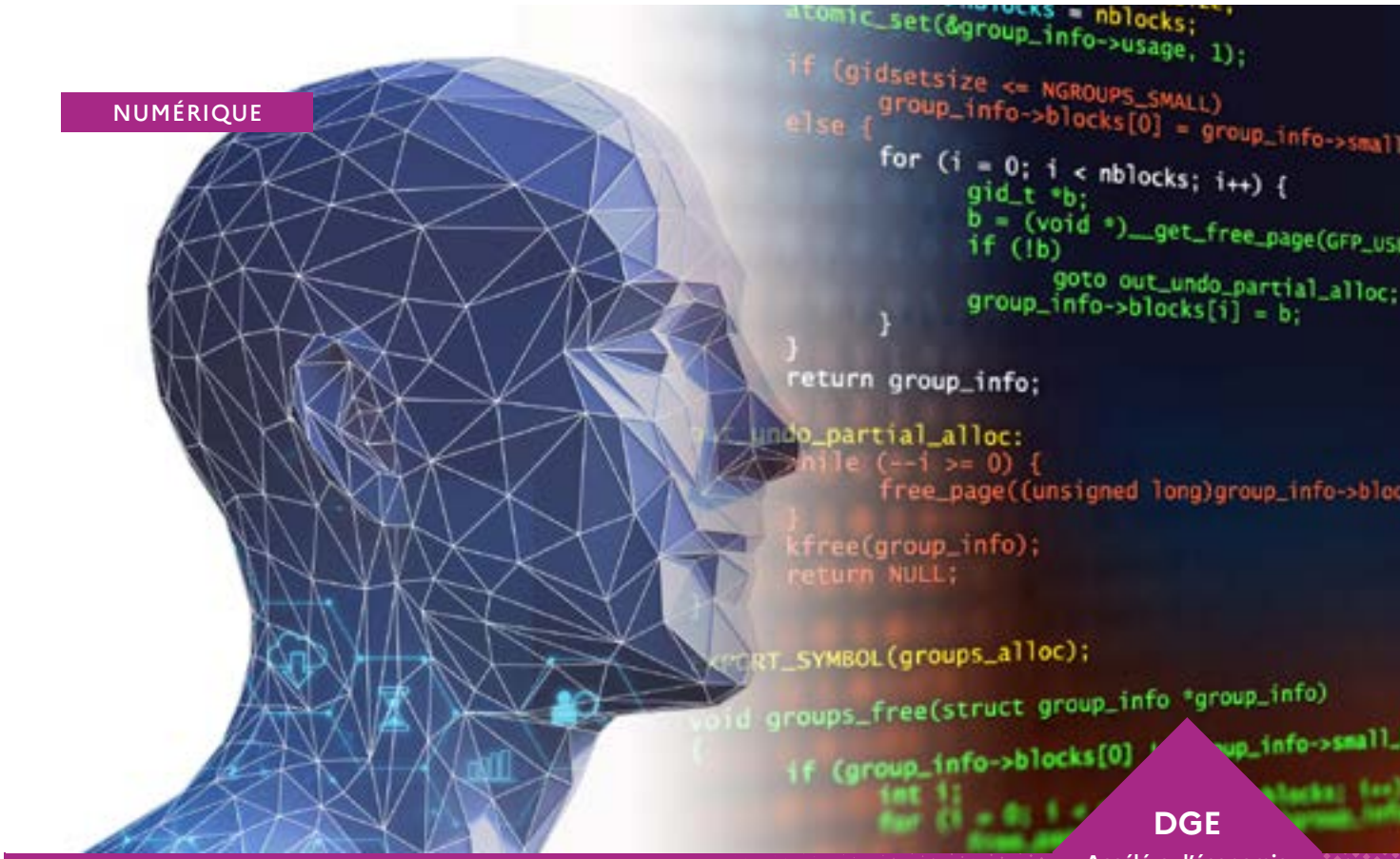
MINISTÈRE
DE L'ÉCONOMIE,
DES FINANCES
ET DE L'INDUSTRIE

Liberté
Égalité
Fraternité

Direction générale
des Entreprises



DGE
Accélérer l'économie
de demain !

Guide de la génération augmentée par récupération (RAG)

Novembre 2024

→ www.entreprises.gouv.fr

Ce premier guide sur l'intelligence artificielle (IA), à destination des entreprises de toute taille, participe aux actions d'accompagnement des entreprises vers le sujet complexe de l'adoption de l'IA. Cette première version du guide est destinée à être enrichie par vos retours et mise à jour des nouvelles informations disponibles. Pour toute remarque ou tout commentaire, n'hésitez pas à nous écrire à ia.dge@finances.gouv.fr.

Thomas Courbe

Directeur général
des entreprises

L'IA générative peut apporter de formidables gains de productivité donc de compétitivité aux entreprises françaises, quelque soient leurs taille, secteur d'activité ou compétences en informatique. Pourtant, le manque de visibilité sur ses apports concrets et sur l'offre disponible ralentit le rythme d'adoption de l'IA par les PME et ETI.

Outre le soutien au développement de l'offre mené avec constance depuis le lancement de la stratégie nationale pour l'IA en 2018, la direction générale des entreprises (DGE) a lancé un plan d'adoption de l'IA dans l'économie. Celui-ci cherche à favoriser la collaboration entre l'écosystème d'IA et les entreprises désireuses d'adopter l'IA, à travers des dispositifs tels que l'appel à projets « Accélérer l'usage de l'IA générative dans l'économie » du plan France 2030, à prodiguer un accompagnement technique à l'adoption de l'IA pour les PME et ETI à travers le programme « IA Booster », mais aussi à pallier le manque d'information sur l'adoption de l'IA par le référencement public des cas d'usages de l'IA. C'est également le sens du présent guide pour l'adoption de la génération augmentée par récupération.

La « génération augmentée de récupération » – aussi appelée RAG (pour retrieval augmented generation) – est une technologie qui consiste à connecter des modèles d'IA générative à des bases de données internes à l'entreprise. Cette technique assure que les modèles d'IA générative connaissent les spécificités de l'entreprise et de son secteur d'activité, ce qui est décisif pour que l'apport de l'IA soit réel et durable, et ce à moindre coût et avec très peu de prérequis techniques. C'est pourquoi la DGE a choisi de mettre en lumière cette technologie, qui est un premier pas pour l'adoption de l'IA générative dans le quotidien des entreprises.

À cet effet, les équipes de la DGE ont mené une consultation auprès d'entreprises développeuses d'IA générative, d'intégrateurs, d'entreprises ayant déjà intégré l'IA générative à leur activité ainsi que de PME et ETI désireuses d'adopter l'IA et a élaboré ce guide d'adoption du RAG.

Destiné aux dirigeants de toutes les entreprises, le présent guide méthodologique expose les cas d'usage les plus prometteurs du RAG et traite avec pédagogie des prérequis, des coûts, des bonnes pratiques et des divers choix technologiques qui se présentent pour mener à bien un projet de déploiement d'une solution de RAG dans une entreprise. Pour s'engager dans l'adoption de l'IA générative au service des problématiques de votre métier, nous sommes à vos côtés. Bonne lecture !

Tirer profit de l'IA générative en y connectant ses données d'entreprise

L'IA générative peut offrir des gains de temps et de productivité considérables en assistant les employés dans la réalisation de tâches d'ordre général (rédaction d'e-mails, de comptes-rendus de réunion) ou de tâches nécessitant une recherche d'informations (par exemple, dans des documents juridiques, des textes de loi, des contrats, de la documentation technique).

Cependant, les grands modèles de langage, aussi appelés large language models en anglais ou LLM (Alfred, Mistral 7B et Large, GPT, LLaMA, etc.) sont entraînés sur des corpus de texte généralistes, ils ne sont donc pas informés du contexte spécifique de votre entreprise et les données sur lesquelles ils s'appuient sont figées lors de leur entraînement. Par conséquent, seuls, ils ne seront pas en mesure de vous donner accès à ces gains de temps et de productivité pour les tâches nécessitant un accès à la base de connaissance de votre entreprise. **Pour en tirer pleinement parti, il est nécessaire de connecter un modèle de langage à vos données d'entreprise.**

La « génération augmentée de récupération » (en anglais « retrieval-augmented generation », RAG) est **une technique qui consiste à améliorer ou augmenter les réponses des modèles d'IA générative, en les alimentant avec des connaissances issues des bases de données de votre entreprise**. Autrement dit, elle permet à un LLM de « consulter » vos données d'entreprise (en temps réel) avant de fournir une réponse. Outre qu'elle permet de contextualiser les modèles génératifs, le RAG permet d'améliorer la traçabilité de l'information et de réduire le risque d'hallucinations (réponses absurdes) par rapport à la simple utilisation d'un modèle d'IA générative.

Cette solution est accessible à toutes les entreprises, quels que soient leurs taille et secteur d'activité. En effet, elle ne nécessite pas de compétences en IA ni même en informatique en interne pour être adoptée, et les données utilisées pour la mettre en place peuvent aussi bien être des e-mails, documents textuels et PDF du quotidien que des bases de données juridiques ou techniques d'une grande complexité.

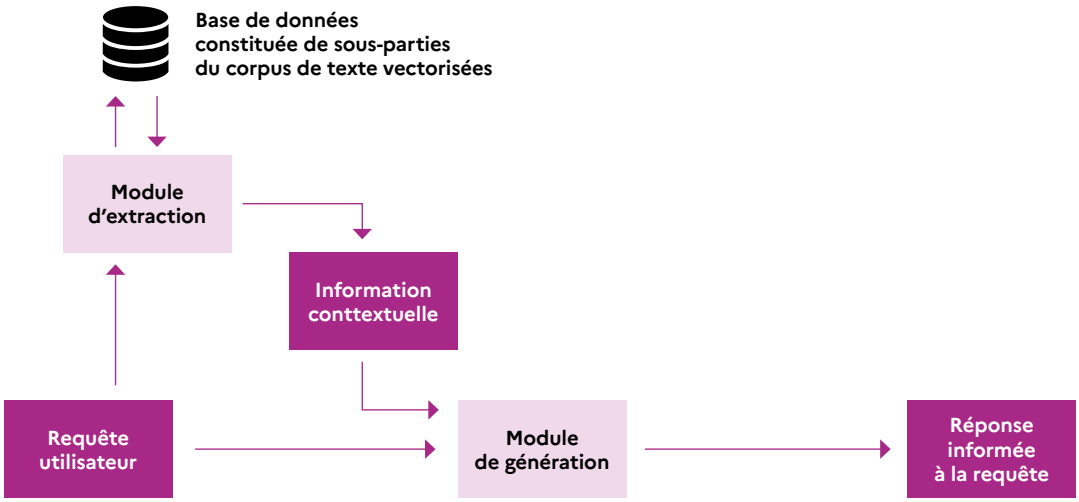
Ce document a pour but de vous guider dans la mise en place d'un outil de type « RAG » dans votre entreprise. Notons que certains points techniques y sont mentionnés afin de donner une compréhension générale de la technologie mais vous n'aurez pas à vous en soucier par vous-mêmes si vous choisissez par exemple d'utiliser un logiciel commercial existant pour déployer votre solution.

L'utilisation de l'intelligence artificielle (IA) générative sur des bases de données d'entreprise permet des gains de productivité importants. Outre ces gains de productivité, le RAG peut améliorer la qualité du service rendu par les employés, par exemple en leur donnant accès à des informations plus complètes, des recommandations plus ciblées ou encore des idées plus novatrices que celles dont ils disposeraient sinon ou en leur permettant de se former.

Néanmoins, afin de tirer parti du potentiel du RAG, il est souhaitable de sélectionner soigneusement les cas d'usage afin de s'assurer que le RAG est pertinent pour y répondre (et qu'un système de recherche plus classique ne peut pas y répondre aussi bien) et d'associer le plus tôt possible les utilisateurs finaux afin de s'assurer que le système répond bien à leur besoin. Ce guide propose un support pour mener ce travail préalable à la mise en place d'un système de RAG.

Comment fonctionne le RAG ?

Figure 1 — Schéma simplifié du fonctionnement d'un système de RAG



Un système de RAG se décompose en deux modules : un module de récupération d'informations et un module de génération :

- → **Le module de récupération d'informations** permet, à partir d'une requête, de chercher dans un corpus documentaire les documents ou extraits de documents les plus pertinents en réponse à cette requête.
- → **Le module de génération** s'appuie sur un grand modèle de fondation (le plus souvent un grand modèle de langue ou LLM) pré-entraîné auquel on a appris à répondre à la combinaison d'une requête utilisateur et d'une information contextuelle (un document ou extrait de document).

Une fois une solution de RAG mise en place, lorsqu'un utilisateur effectue une requête, le module de récupération identifie au sein de la base documentaire les documents ou extraits de documents les plus pertinents au vu de cette requête. Ensuite, le module de génération utilise un modèle d'IA générative pour générer une réponse à la requête en tenant compte des documents ou passages de documents identifiés par le module de récupération d'informations.

L'essentiel sur le RAG

Quels sont les bénéfices du RAG ?

Les bénéfices attendus du RAG sont les suivants :

- **Gains de temps ou de productivité** : un système de RAG peut proposer aux employés des e-mails prérédigés, des comptes-rendus de réunion prérédigés ou chercher à leur place l'information qui leur est nécessaire dans une base documentaire de grande ampleur, ce qui peut leur faire gagner un temps considérable ;
- **Amélioration de la qualité du travail effectué** : pour toute tâche nécessitant de la recherche d'informations préalable dans une base de connaissances particulièrement vaste, le système de RAG pourra être plus exhaustif qu'une recherche manuelle, ce qui peut améliorer le niveau d'information dont disposera l'employé utilisateur du système de RAG donc la qualité de son travail ;
- **Formation** : un système de RAG peut également constituer un outil de formation interne ou de formalisation et de partage des connaissances susceptible de répondre aux questions des employés pour les faire progresser.

Quels cas d'usage ?

Le RAG est approprié pour effectuer des tâches qui nécessitent d'interagir avec une base de connaissances interne à l'entreprise et requièrent des réponses rédigées en langage naturel. Afin d'illustrer le propos, voici ci-dessous quelques cas d'usage récurrents du RAG ainsi que la base documentaire associée à chacun :

Cas d'usage	Base documentaire
Assistant d'entreprise généraliste : rédaction de comptes-rendus de réunions, synthèse et rédaction d'e-mails, gestion du planning	E-mails, plannings, enregistrement audio de réunions
Assistant juridique d'entreprise : vérification de la conformité et aide à la rédac-tion de contrats	Contrats passés par l'entreprise
Assistant RH : aide à la rédaction de documents RH ou contrats, recherche dans les données RH internes	Conventions collectives et documents RH internes (jours de congé…), fiches de postes et CV
Assistant de recherche dans de la docu-mentation technique : métiers de la main-tenance en industrie, bureaux d'études	Documentation technique (ex : nomenclature et description des opéra-tions de maintenance)
Assistant à la création de nouveaux produits	Historique des produits créés, de leurs descrip-tions, de l'évolution de ces produits et des échanges les concernant
Assistant commercial	Description détaillée de chaque produit de l'offre commerciale de l'entreprise

Dans chaque cas, l'utilisateur pourra interroger la base documentaire et obtenir des réponses organisées. Le LLM joue le rôle d'interlocuteur en langage naturel, la base documentaire de source d'informations à laquelle l'interlocuteur à accès.

Mode d'intégration du RAG		Points forts	Points faibles
Solutions clés en main	**Logiciel sous licence** Le système de RAG est acquis sous forme de licence logicielle et installé sur les serveurs de l'entreprise	Ne nécessite pas de compétences en interne Déploiement rapide Maîtrise des données	Personnalisation limitée (pas toujours de logiciel existant adapté à des cas d'usage spécifiques) Coût de maintenance
	Software as a Service (SaaS) Solution hébergée dans le cloud avec abonnement	Ne nécessite pas de compétences en interne Déploiement rapide Hébergement, maintenance et mises à jour pris en charge	Pas toujours de SaaS existant adapté à des cas d'usage spécifiques Requiert que les données sortent de l'entreprise
Solutions sur mesure	**Développement par un intégrateur** Système personnalisé installé par un prestataire externe	Système sur-mesure donc bien adapté aux spécificités du cas d'usage	Nécessite un minimum de compétences en informatique pour dialoguer avec l'intégrateur Coût de maintenance
	Développement en interne Système développé par les équipes internes de l'entreprise	Système sur-mesure donc bien adapté aux spécificités du cas d'usage Contrôle total sur les choix technologiques, la personnalisation et la maintenance	Nécessite des compétences en interne (développement, IA, administration système) Déploiement plus long Nécessite de mobiliser des équipes pour la maintenance

Quel coût en moyenne ?

Les coûts dépendent fortement du volume d'utilisation.

Un produit SaaS a un coût d'abonnement annuel très variable selon les cas d'usage et les fonctionnalités proposées.

À partir d'entretiens réalisés avec des professionnels du secteur, la prestation d'installation d'un système par un intégrateur est estimée à environ à 100 000 €, en ordre de grandeur.

En cas de développement en interne, le coût comprendra la main d'œuvre affectée (au moins un data scientist, un développeur et un administrateur-système mobilisés pendant au moins 3 mois).

Hormis pour le SaaS, il faut ajouter les coûts de la puissance de calcul pour le mode d'hébergement retenu :

→ En cas d'hébergement en local (on-premise), le coût de la puissance de calcul couvrira l'achat des processeurs GPU (coût fixe) et le coût de l'électricité (coût variable). En supposant qu'on utilise le modèle LLaMa-8B-4bits, le coût d'achat du GPU sera de 3250 € et le coût de l'électricité sera de 0.10€ pour 1000 requêtes. En supposant que chaque employé effectue 300 requêtes par semaine, le coût de l'électricité reviendra donc à environ 2 € par employé utilisateur et par an.

→ En cas d'hébergement sur le cloud, le coût de la puissance de calcul sera entièrement un coût variable. En supposant qu'on utilise le modèle Mistral Small, il sera de 7.8€ pour 1000 requêtes. En supposant que chaque employé effectue 300 requêtes par semaine, le coût du cloud reviendrait donc à environ 106 € par employé utilisateur et par an.

L'IA générative, à travers le RAG, peut automatiser ou assister les tâches qui nécessitent une recherche dans une base documentaire (contrats, e-mails, documentation technique…). Une fois une telle tâche identifiée au sein de votre entreprise, quelles questions se poser pour savoir si le RAG est pertinent pour y répondre ?

- Cette tâche met-elle en jeu des bases documentaires de grande ampleur, internes à l'entreprise ?
- Ces bases documentaires changent-elles souvent (ajouts ou modifications de documents) ?
- Ces bases documentaires sont-elles organisées en une base de données structurée ?
- Les résultats de la recherche documentaire doivent-ils être intégrés à une réponse rédigée en langage naturel ?
- Cette tâche nécessite-t-elle une traçabilité de l'information (pouvoir citer les références des documents de la base documentaire employés pour répondre) ?

Pour les entreprises ne disposant pas de bases de données bien organisées ou de grande ampleur les cas d'usage simples tels que la rédaction automatique d'e-mails peuvent apporter des gains de productivité importants grâce à un système de RAG basé sur le serveur de courriels de l'entreprise.

De tels cas d'usages peuvent servir de première adoption du RAG dans l'entreprise, de façon à acculturer les employés et de mesurer les premiers retours sur investissement, avant possiblement d'installer des systèmes plus sophistiqués basés sur des données plus complexes, avec des gains de productivité plus importants.

Dans quels cas le RAG est-il moins pertinent ?

En général, le RAG est moins pertinent pour des cas d'usages qui dépassent la recherche d'information, par exemple lorsqu'il s'agit de formuler des recommandations ou de faire des comparaisons entre documents. Pour ces derniers cas, on peut utiliser un système de RAG qui comporte des briques technologiques explicitement développées dans le but de s'assurer de la qualité des recommandations formulées par le système de RAG.

Aujourd'hui, le RAG est en général plus performant sur des données de texte que sur des modalités autres, notamment les tableaux. De plus, le RAG est moins adapté aux requêtes qui nécessitent d'avoir accès à l'ensemble des documents de la base documentaire plutôt qu'à un petit nombre d'entre eux (par exemple, un résumé ou une synthèse comparative entre un grand nombre de documents). Cependant, le RAG est en évolution rapide et les méthodes de RAG les plus récentes tendent à pallier ces deux faiblesses.

Quels sont les points de vigilance pour assurer la bonne appropriation de l'IA générative au sein de l'entreprise ?

L'IA générative en général et les systèmes de RAG en particulier peuvent susciter de défiance ou scepticisme. Il importe, pour les prévenir, de veiller à coconstruire les solutions avec les futurs utilisateurs.

a. Implication des utilisateurs

L'implication des utilisateurs peut avoir lieu à différents moments du cycle de mise en place du RAG :

- Les futurs utilisateurs doivent être associés à la définition des cas d'usage les plus pertinents

— Ils peuvent ensuite contribuer à élaborer les questions-types et réponses-types qui seront fournies au système de RAG lors de sa mise en place

— Ils peuvent ensuite participer aux points d'étape avec les équipes techniques (internes ou externes selon le mode de déploiement retenu) au cours de la phase d'expérimentation

— Il est souhaitable de développer l'interface utilisateur de manière à permettre à l'utilisateur de faire des retours et d'itérer sur les réponses du système de RAG de manière à les affiner.

b. Acculturation à l'IA

Il est également nécessaire de mener un travail d'acculturation à l'IA auprès des équipes. Ce travail d'acculturation comporte deux volets :

— Former à l'utilisation de l'outil et notamment au prompt engineering, c'est-à-dire à la manière de formuler des requêtes au système de RAG pour obtenir les réponses les plus satisfaisantes possibles. En effet, les modèles d'IA sont très sensibles à la manière dont est formulée la demande, aux termes employés et à l'ordre dans lequel les requêtes sont formulées. De nombreux guides didactiques de prompt engineering sont disponibles en ligne.

— Démystifier l'IA :

  – Donner une vision objective des performances des modèles d'IA, des bénéfices qu'ils apporteront (gain de temps, exhaustivité) comme de leurs limites de manière à éviter le double écueil de la perception d'une IA omnipotente, qui conduit nécessairement à la déception, et d'une IA inutile

  – Souligner que l'IA vise à assister les métiers dans leurs tâches et en aucun cas à les remplacer

Une fois un système de RAG installé et mis en service, il est important d'en évaluer les performances objectivement grâce à des outils dédiés et d'en faire la maintenance de façon appropriée. L'évaluation objective de la performance du système est un facteur déterminant dans l'adoption de l'outil par les employés, et les briques technologiques correspondantes existent.

c. Considérations éthiques et de sécurité

— **Protection des données**

La protection des données sensibles des entreprises est un enjeu crucial. Aussi est-il primordial de s'assurer que les risques de fuite de données via le système de RAG restent maîtrisés. Plusieurs critères permettent d'évaluer le risque de fuite de données :

  – La maîtrise de l'infrastructure utilisée pour le pré-entraînement et l'utilisation du système de RAG : l'installation du système de RAG on-premise est par construction l'architecture qui présente le meilleur degré de maîtrise, à condition de disposer, y compris dans la durée, des compétences pour la maintenir à l'état de l'art. Le recours à des services de cloud labellisés SecNumCloud est une alternative permettant de bénéficier des avantages et de la flexibilité des infrastructures cloud sans transiger sur le niveau de protection des données face à des attaques informatiques ou via l'application de lois extraterritoriales ;

  – La maîtrise du modèle/système d'IA : le recours à des modèles de confiance permet de limiter les fuites de données via le modèle d'IA. Le recours à un modèle open source présente par construction une meilleure auditabilité par des tiers.

En somme, c'est l'ensemble de l'architecture (infrastructure, modèles, modalités d'intégration, type de données manipulées) qui permettra à l'entreprise de déterminer si les risques de fuite de données apparaissent correctement couverts. Lorsque le développement du système de RAG est confié à un intégrateur, la protection contre les fuites de données devrait être intégrée aux cahiers des charges.

— **Conformité au règlement IA européen (RIA ou AI Act)**

Le règlement sur l'intelligence artificielle est entré en vigueur le 1er août 2024. Ce texte vient encadrer certains usages des modèles et systèmes d'IA.

Pour le cas du RAG, des obligations de transparence[1] pourraient s'appliquer à compter du 2 août 2026. Par exemple, les systèmes d'IA destinés à interagir avec des personnes physiques doivent être conçus et développés de manière à ce que les personnes physiques concernées soient informées qu'elles interagissent avec un système d'IA.

Certains systèmes d'IA sont par ailleurs considérés comme des systèmes à haut-risque par le règlement et seront soumis à partir du 2 août 2026 à des obligations supplémentaires de mise en conformité [2] . Par exemple, un système de RAG destiné à être utilisé pour le recrutement ou la sélection de personne pourrait être considéré comme un système d'IA à haut-risque et devrait donc respecter les exigences de conformité qui sont attachées à cette catégorie par le règlement IA.

La réalisation d'une analyse détaillée de l'usage du système de RAG, au cas-par-cas et au préalable de son déploiement, sera nécessaire afin de déterminer si ce système d'IA est concerné par le règlement sur l'intelligence artificielle. Auquel cas, il faudra prévoir une mise en conformité du système d'IA aux exigences du règlement.

1 – Il s'agit des systèmes listés à l'article 50 du règlement (UE) 2024/1689 établissant des règles harmonisées concernant l'intelligence artificielle.
2 – Il s'agit des systèmes couverts par l'article 6 du règlement (UE) 2024/1689 établissant des règles. harmonisées concernant l'intelligence artificielle et notamment ceux répertoriés à l'annexe III. A noter que ce même article prévoit des dérogations à cette classification à haut risque dans certains cas. Des lignes directrices sont prévues par le règlement qui qui permettront de préciser la liste des systèmes d'IA qui sont effectivement considérés comme étant à haut risque et ceux qui ne le sont pas.

Guide détaillé pour l'adoption du RAG

1. Choix du mode d'intégration du système de RAG

Les modes d'intégration du RAG sont les suivants :

Solutions clés en main	**Logiciel sous licence** Le système de RAG est acquis sous forme de licence logicielle et installé sur les serveurs de l'entreprise
	Software as a Service (SaaS) Solution hébergée dans le cloud avec abonnement
Solutions sur mesure	**Développement par un intégrateur** Système personnalisé installé par un prestataire externe
	Développement en interne Système développé par les équipes internes de l'entreprise

Questions à se poser pour déterminer le mode d'intégration le plus adapté :

- **Compétences internes**

  Si l'on n'a aucune compétence en informatique en interne, il conviendra de s'orienter vers un produit SaaS.

  Si l'on dispose d'au moins un data scientist, un développeur et un administrateur système, toutes les pistes sont envisageables : on peut développer un système de RAG en interne, faire appel à un intégrateur ou utiliser un produit SaaS. En effet, un projet de RAG requiert un data scientist pour s'assurer de l'adéquation entre les données et les méthodes utilisées, et vectoriser la base de données, un administrateur système pour la configuration et la maintenance du matériel, la sécurité et l'informatique et le déploiement de l'outil de RAG et un développeur pour la création de l'outil de RAG.

  Si l'on ne dispose pas des compétences ci-dessus mais que l'on dispose néanmoins de compétences génériques en informatique, alors mieux vaut utiliser un produit SaaS ou faire appel à un intégrateur externe pour installer un système de RAG. En cas d'appel à l'intégrateur, les développeurs pourront collaborer avec l'intégrateur, ce qui constitue aussi un vecteur de montée en compétences des ingénieurs de l'entreprise.

- **Moyens financiers**

  En général, un produit SaaS coûte plus cher qu'une prestation d'un intégrateur qui elle-même est plus onéreuse qu'un développement en interne. En effet, l'ordre de grandeur d'un produit SaaS, selon la gamme choisie, peut être de 10 000 €/an ou de 100 000 €/an.

  Une prestation d'un intégrateur est de l'ordre de 100 000 € (payés une seule fois), un déploiement en interne coûtera le coût de la main d'œuvre affectée

à ce projet (au moins un data scientist, un développeur et un administrateur système pendant plus de 3 mois).

Hormis en cas d'utilisation d'un produit SaaS, le coût de l'infrastructure d'hébergement (cloud ou en local) s'ajoutera au coût de la solution elle-même.

- **Complexité et spécificité du cas d'usage**

  Si le cas d'usage, via son vocabulaire ou ses concepts manipulés, est complexe ou spécifique, il est recommandé de développer un système de RAG en interne ou de faire appel à un intégrateur. Dans ce cas, des étapes supplémentaires de pré-traitement de la donnée (par exemple, de création de graphes de connaissances) pourront prendre en compte cette complexité et améliorer la qualité du système.

  À l'inverse, si le cas d'usage est relativement simple et courant et ne met pas en jeu de vocabulaire trop spécifique, on pourra s'appuyer sur un produit SaaS.

  Par exemple, un assistant pour la conception de nouveaux médicaments mettra en jeu du vocabulaire spécifique (chimie, pharmacie…) et complexe donc on privilégiera un appel à intégrateur ou un développement en interne. À l'inverse, un assistant pour rédiger des comptes-rendus de réunions et des e-mails destiné à toutes les fonctions d'entreprise sera moins complexe et plus générique donc on pourra utiliser un produit SaaS.

- **Délai de mise en place souhaité**

  Pour mettre en place une solution de RAG personnalisée en faisant appel à un intégrateur, il faudra compter environ 3 mois. Pour un développement en interne, 3 mois seront nécessaires pour un produit minimum viable et quelques mois supplémentaires seront requis pour un outil fonctionnel[1]. Si on souhaite un déploiement plus rapide, une solution SaaS sera plus adaptée.

— **Besoin de simplicité de mise à jour**

  Si l'on utilise un SaaS, le RAG sera mis à jour automatiquement au gré des mises à jour du logiciel, sans coût autre que celui de l'abonnement annuel ou mensuel. C'est donc l'option qui garantit la mise à jour la plus facile et la moins coûteuse.

  Si l'on fait appel à un intégrateur, la maintenance du RAG pourra être incluse dans le coût initial facturé par l'intégrateur ou engendrer un coût supplémentaire.

  Si l'on effectue du RAG en interne, la mise à jour nécessitera de mobiliser ses équipes techniques.

1 – Toutefois, les durées de déploiement indiquées ci-dessus le sont à titre indicatif et la durée de déploiement dépend de la complexité du cas d'usage retenu.

2. Choix de l'hébergement du système de RAG

Hormis recours à un produit SaaS, les options possibles pour l'hébergement du système sont les suivantes[2] :

Questions à se poser afin de choisir un mode d'hébergement :

Mode d'hébergement		Points forts	Points faibles
Serveurs internes		Contrôle total sur les données et l'infrastructure	Besoin de ressources pour l'achat et la maintenance des équipements, notamment des GPU *(Graphical Processing Units)*
Hébergement dans le cloud	***Cloud privé*** Hébergement chez un fournisseur de cloud sur des serveurs dédiés à votre entreprise	Équilibre entre contrôle et flexibilité Garanties de sécurité élevées	
	Cloud public Hébergement chez un fournisseur de cloud sur des serveurs partagés	Flexibilité du service	Les données transitent entre l'entreprise et le fournisseur cloud
Software as a service (SaaS) Le fournisseur gère l'hébergement, la maintenance et les mises à jour		Pas de compétences nécessaires au maintien d'un cloud ou de GPU	Peu de contrôle sur les données, cependant les services certifiés SecNumCloud offrent des garanties de sécurité

— **Vitesse de déploiement souhaitée**

Un déploiement on-premise est plus long qu'un déploiement sur le cloud (à titre indicatif, un même projet peut nécessiter un déploiement de 6 mois on-premise contre 2 mois dans le cloud).

— **Exigence en matière de sécurité des données**

Un hébergement on-premise comprend par construction les meilleures garanties en termes d'isolement des données, ce qui constitue une garantie de sécurité des données si la cybersécurité des systèmes d'information est bien assurée.

En cas d'hébergement dans le cloud, les données transitent entre l'entreprise

2 – Notons toutefois que l'hébergement peut être différencié selon les briques technologiques constitutives du RAG (cf. section 3 pour le détail des différentes briques technologiques). Ainsi, le cas de déploiement 2 (cf. section 8) repose sur une brique de vectorisation et de recherche vectorielle on-premise et un appel au LLM dans le cloud.

cliente et les serveurs du fournisseur de cloud. Cependant, certains services de cloud offrent des garanties fortes en matière de sécurité des données. Ainsi, la certification SecNumCloud délivrée par l'ANSSI en France est l'une des plus exigeantes au monde en matière de sécurité des données. Certaines solutions de cloud ou de SaaS pour le RAG sont labelisées SecNumCloud.

— **Nombre de requêtes attendu**

Si le nombre de requêtes effectuées auprès du système de RAG est élevé, il peut être financièrement plus rentable d'utiliser un système de RAG on-premise.

En effet, l'hébergement dans le cloud entraîne uniquement des dépenses de fonctionnement liées à l'utilisation du cloud (par exemple, pour une PME avec des besoins occasionnels utilisant Mistral Small, cela représentera 7.8€ pour 1000 requêtes). Inversement, un hébergement en local entraîne des dépenses d'investissement pour acheter les GPU (par exemple, 3250 € pour un GPU Nvidia RTX 4090 pour une PME avec des besoins occasionnels utilisant le LLM LLaMA-8B-4 bits) et des dépenses d'électricité (dans l'exemple, 0.1€ pour 1000 requêtes) [3].

— **Compétences internes**

Le RAG repose le plus souvent sur de grands modèles de langage (LLM). Ces modèles d'IA possèdent un grand nombre de paramètres (plusieurs milliards). Par conséquent, l'appel à des modèles de ce type (inférence) requiert une puissance de calcul considérable que seules peuvent fournir des puces appelées Graphical processing units (GPU). En cas de déploiement en interne, il faut donc disposer de serveurs munis de GPU. Etant donné que les GPU, une fois installés, doivent être maintenus, les compétences nécessaires au maintien des GPU sont nécessaires à l'installation d'une solution de RAG en local.

3. Prétraitement des données pour le RAG

Le prétraitement des données est une étape essentielle pour garantir la performance du système de RAG. Il s'agit de préparer, structurer et enrichir les données afin qu'elles soient exploitées de manière optimale.

Quelles sont les étapes de prétraitement et de structuration des données ?

Le présent guide fournit à titre indicatif la liste des étapes du pré-traitement de données pour la mise en place d'un système de RAG. Cette partie plus technique vise à poser le vocabulaire pour une discussion avec un éventuel prestataire de systèmes de RAG, mais il n'est pas attendu de connaître les détails techniques de ces étapes.

1) Préparation des données

— **Conversion des données** : conversion des documents dans une unique modalité [4], généralement le texte. Cela peut inclure l'utilisation de la reconnaissance optique

3 – Tous les coûts cités dans ce paragraphe ont été évalués en juin 2024.
4 – Une modalité de données désigne un type de données tel que le texte, l'image, la vidéo, la série temporelle.

de caractères (OCR) pour convertir des documents scannés en texte ou la conversion de présentations en texte.

— Nettoyage et normalisation des données : suppression des doublons, correction des erreurs, mise en cohérence des données (supprimer les documents donnant des informations contradictoires), retrait d'éventuels en-tête, pieds de page, pages de garde, tables des matières, anonymisation ou pseudonymisation des données à caractère personnel

2) Structuration des données

— **Segmentation (ou chunking)** : division des documents en segments élémentaires. Lors de cette étape, un choix doit être effectué quant à la méthode de chunking et la taille des segments, qui est un paramètre crucial dans la réussite du RAG et nécessite une bonne interaction entre experts techniques (internes ou externes à l'entreprise utilisatrice) et experts du métier visé (internes à l'entreprise utilisatrice, qui peuvent être les futurs usagers du système de RAG).

— **Construction d'un graphe de connaissances (en option)** afin de matérialiser les liens entre les différents termes et concepts propres à l'entreprise (celui-ci peut s'appuyer sur des nomenclatures ou ontologies déjà présentes au sein de l'entreprise)

— **Enrichissement des segments avec des métadonnées** : une fois les segments (ou chunks) découpés, il peut être judicieux d'y ajouter des métadonnées pour fournir au système des informations de contexte sur les segments : paragraphe, chapitre d'appartenance, date d'écriture, notions sur lesquelles porte le document… [5]

3) Construction de la base de données vectorielle

— **Vectorisation** : conversion de chaque segment de document en un vecteur, c'est-à-dire en une séquence de nombres. Elle peut reposer sur des méthodes statistiques classiques, sur des réseaux de neurones ou encore reposer sur une approche hybride.

— **Indexation de la base de données vectorielle** : il s'agit d'organiser les vecteurs de la base de données afin d'en faciliter l'accès lors de la phase de récupération d'informations.

Qui est responsable du prétraitement des données ?

— **Développement en interne** : Si le RAG est développé en interne, le prétraitement des données est à la charge de l'entreprise utilisatrice.

— **Développement par un intégrateur** : dans ce cas, le prétraitement des données peut être à la charge du client ou de l'intégrateur, auquel cas ce dernier mobilisera les équipes du client pour leur expertise métier, indispensable au cours du processus de prétraitement des données.

— **Solutions clés en main (SaaS ou logiciel sous licence)** : Les produits SaaS intègrent souvent des outils de prétraitement automatisés, réduisant la charge pour votre entreprise.

5 – L'ajout de métadonnées permet une phase de recherche d'informations plus ciblée. Ainsi, selon la question posée, la recherche d'informations pourra se focaliser sur les documents issus de certaines sources, de certaines dates ou périodes ou ceux abordant certains sujets. Il permet aussi de donner à l'utilisateur davantage de détails sur les sources utilisées pour produire la réponse.

Le prétraitement des données constitue-il un investissement durable ?

Oui, le prétraitement des données constitue un investissement durable qui profite à l'ensemble de votre entreprise. Il facilite non seulement le déploiement du RAG, mais également d'autres projets de valorisation des données, tels que la recherche documentaire avancée ou d'autres applications d'IA.

Il est pertinent de procéder progressivement à la mise en place d'une base de données structurée dans son entreprise, en commençant par les données plus génériques (par exemple, les courriels) qui permettent une première adoption du RAG. Cette première adoption permet d'amorcer une acculturation des employés et la mesure d'un premier retour sur investissement. Ensuite, on peut aborder des cas d'usages plus complexes (par exemple, avec des données financières ou juridiques), avec des gains de productivité plus importants, et monter en puissance progressivement l'usage de l'IA dans sa structure.

Le prétraitement des données doit-il être reproduit à chaque arrivée de nouveaux documents ?

Non, car les modules de prétraitement de données sont automatisés : une fois ceux-ci développés, on fait en sorte que chaque nouveau document ajouté à la base documentaire passe par ce module de prétraitement.

4. Quelles sont les briques technologiques du RAG ?

Le tableau suivant recense les briques technologiques constitutives d'une solution de RAG (voir aussi l'encadré « Comment fonctionne le RAG » et l'Annexe A).

Brique technologique	Description
Prétraitement de la donnée	Les étapes suivantes peuvent ou non être requises selon le cas d'usage et le modèle utilisé (voir aussi ci-dessus) : – Conversion des données initiales en une modalité unique –Nettoyage des données – Graphe de connaissances – Anonymisation – Segmentation – Enrichissement – Vectorisation – Indexation
Module de recherche	Module permettant d'interroger la base de données pour en extraire les segments de documents les plus pertinents en réponse à une requête d'utilisateur **Trois familles de méthodes de recherche existent :** → **Les méthodes de recherche lexicale** (recherche des segments de documents présentant le plus de mots-clés communs avec la requête) **Limites** : Synonymes et sens multiples mal pris en charge → **Les méthodes de recherche sémantique** (recherche des segments de documents dont la représentation vectorielle ressemble le plus à la représentation vectorielle de la requête, autrement dit dont le sens est le plus proche du sens de la requête) **Limites** : plus cher à déployer et moins performant en cas d'adéquation exacte entre requête et documents → **Les méthodes de recherche hybrides** visent à combiner le meilleur des méthodes de recherche lexicales et sémantiques
Grand modèle de langage (LLM)	Modèle d'IA générative qui produit des réponses en langage naturel à une requête de l'utilisateur en s'appuyant sur les données récupérées par le module de recherche
Orchestration de LLM (seulement pour les systèmes de RAG multi-agents)	Coordination de plusieurs modèles ou agents pour optimiser les performances (par exemple, interroger différents LLMs en fonction de la requête)
Évaluation du RAG	Outils et méthodes pour mesurer la qualité et la pertinence des réponses su système de RAG lors de son fonctionnement
Interface utilisateur	Application ou interface par laquelle les utilisateurs interagissent avec le système de RAG

5. Évaluation et maintenance du système de RAG

Évaluation du système de RAG

L'évaluation régulière du système de RAG est essentielle pour garantir sa performance et sa fiabilité. Elle permet de détecter et de corriger les « hallucinations » (réponses erronées), d'assurer la pertinence des résultats et de maintenir la confiance des utilisateurs.

Des briques technologiques dédiées mettent en œuvre différentes méthodes d'évaluation de systèmes de RAG déployés :

- Retours des utilisateurs : Recueillir les retours (quantitatifs ou qualitatifs) des utilisateurs finaux sur la qualité des réponses.
- Mesures quantitatives : Utiliser des indicateurs tels que le taux d'hallucination, la pertinence des documents récupérés, le temps de réponse.
- Évaluation par un grand modèle de langage (LLM as a judge) : Utiliser un grand modèle de langage pour évaluer les réponses générées en les comparant à des références.

Maintenance du système de RAG

La maintenance du système de RAG assure sa performance continue malgré l'évolution des données, des besoins et des technologies. Les actions de maintenance doivent suivre étroitement l'évaluation du système en fonctionnement.

Bonnes pratiques de maintenance :

- **Surveillance continue** : Mettre en place des outils de suivi pour détecter les baisses de performance, notamment via l'évaluation en continu du système.
- **Mises à jour régulières** : Actualiser les modèles, intégrer de nouveaux documents et adapter le système aux évolutions du métier. L'intégration d'un grand volume de nouveaux documents dans la base de données peut appeler une nouvelle phase de pré-traitement des données.
- **Gestion des dérives** : Identifier et corriger les dérives dans les performances, par exemple dues à l'augmentation du volume de données.
- **Formation continue** : Sensibiliser les utilisateurs aux nouvelles fonctionnalités et aux bonnes pratiques d'utilisation.

Responsabilité de la maintenance :

- **Solutions sur mesure** : La maintenance peut être assurée par l'intégrateur ou par vos équipes internes, selon le contrat et les compétences disponibles.
- **Solutions clés en main** : Les fournisseurs de SaaS assurent généralement la maintenance et les mises à jour dans le cadre de l'abonnement souscrit.

6. Difficultés en matière de RAG et solutions pour les pallier

Difficultés courantes :

→ **La recherche d'informations peut passer à côté de certains documents pertinents** car la recherche lexicale n'est pas robuste aux variations lexicales telles que l'emploi de synonymes et la recherche sémantique peut passer à côté en raison de variations de formulation (cf. section 4) mais aussi car la recherche de documents se limite souvent au top N de segments de documents les plus proches avec N un nombre fixé

Solutions possibles :

— Méthodes de recherche hybrides entre recherche sémantique et lexicale (cf. section 4)

— Construction d'un graphe de connaissances pour rendre le système plus robuste aux variations de vocabulaire ou de formulation (étape supplémentaire dans le pré-traitement des données)

— Entraînement du système de RAG sur des questions-réponses avec des formulations diversifiées

— LLM orchestrateur ou règles de décision permettant de faire varier le nombre de segments de documents récupérés en fonction de la requête utilisateur

→ **Persistance d'hallucinations : bien que le RAG réduise les hallucinations (réponses aberrantes) des LLM, celles-ci peuvent toujours se manifester**

Solutions possibles :

— Évaluation régulière du système de RAG pour détecter les hallucinations potentielles et amélioration continue du système

— Mise en qualité des données (élimination des contradictions internes de la base documentaire, cf. section 4)

— Construction de graphes de connaissances pour renforcer la cohérence

→ **Performance moindre sur des documents multimodaux (par exemple, un document PDF constitué de texte, d'images et de tableaux)**

Solution possible :

— Adoption de RAG multimodal, intégrant des modèles capables de traiter différentes modalités (texte, image, etc.) par la sélection d'intégrateurs développant ces technologies avancées

7. Innovations en matière de RAG

→ **Modèles multimodaux :** Intégration de modèles vision-langage pour traiter directement des documents complexes sans conversion préalable.

→ **Orchestration :** Utilisation d'un LLM dit « orchestrateur » chargé de répartir les requêtes des utilisateurs entre différents modes de prétraitement de données, modules de recherche d'informations et modèles pour assurer un traitement adapté à chaque requête et améliorer la pertinence et la précision des réponses.

→ **Agentisation du RAG :** développement de LLM spécialisés dits « agents » communiquant entre eux pour améliorer la pertinence et la précision des réponses.

8. Quelques pistes de discussion sur la mise en place d'un système de RAG

On propose ci-dessous quelques thèmes relatifs au RAG qui dépassent les sujets traités dans ce guide, afin de poursuivre la discussion avec un prestataire ou fournisseur de systèmes de RAG.

Confidentialité des données et conformité réglementaire :

- **Protection des données sensibles :** Mettre en place des mesures pour assurer la confidentialité des informations, notamment lors de l'utilisation de services *cloud*.
- **Anonymisation et pseudonymisation :** Supprimer ou masquer les données personnelles et d'une manière générale se conformer aux réglementations (RGPD).
- **Respect des lois en vigueur :** Assurer la conformité avec les réglementations locales et internationales relatives à la protection des données et de la vie privée.
- **Documentation et historique de l'utilisation du RAG :** Maintenir des enregistrements des traitements effectués pour faciliter les audits et les contrôles.

Gestion des biais et équité :

- **Identification des biais :** Être conscient des biais présents dans les données ou les modèles qui pourraient conduire à des réponses injustes ou discriminatoires.
- **Correction des biais :** Mettre en place des processus pour détecter et atténuer les biais dans le système d'évaluation du RAG pour assurer l'équité et l'inclusion.

Transparence et explicabilité :

- **Communication avec les utilisateurs :** Informer les utilisateurs sur le fonctionnement du système, ses limites et les données utilisées.
- **Explicabilité des décisions :** Fournir des explications claires sur les réponses générées, y compris les sources et les références.

Sécurité des systèmes :

- **Cybersécurité :** Protéger le système contre les cyberattaques, les intrusions et les fuites de données.
- **Gestion des accès :** Contrôler les droits d'accès pour limiter l'utilisation du système ou les requêtes à telles bases de données aux personnes autorisées.

CONCLUSION

La mise en place d'un système de RAG peut apporter des gains significatifs en productivité et en qualité de service au sein de votre entreprise. Ce guide vous a présenté les étapes clés, les choix à effectuer et les bonnes pratiques pour réussir cette intégration. La diversité des modes d'intégration fait du RAG une technique accessible à des entreprises de toutes tailles, disposant ou non de compétences internes en IA ou en informatique, disposant de grandes bases de données ou de documents ordinaires, et avec une large gamme de moyens financiers.

Annexe A
Quelques exemples de déploiement de RAG

Cas de déploiement 1

Entreprise de transport

Le cas d'usage : Assistant de maintenance d'infrastructures de transport
Base documentaire associée : Description détaillée de l'ensemble des opérations de maintenance des infrastructures
Mode d'intégration : Appel à un intégrateur
Mode d'hébergement : Cloud
Spécificités techniques : Création de persona (rôles propres à chaque famille de métier), qui permettent au système de RAG de fournir des réponses différenciées selon le métier de l'utilisateur qui formule la requête (cela permet notamment au RAG de chercher dans la base documentaire propre au métier correspondant)
Retour d'expérience : Des problématiques d'acceptabilité ont été rencontrées. En effet, la non-exhaustivité des documents sélectionnés par le RAG pour ses réponses gênaient beaucoup d'employés, une méfiance à l'égard de l'IA s'est manifestée car celle-ci était perçue comme « aléatoire ».

Cas de déploiement 2

Entreprise de la défense

1. Compagnon de programmation (code)

Le cas d'usage : Assistant de programmation qui permet de générer du code dans des langages de programmation peu représentés dans le jeu d'entraînement du LLM mais bien représentés dans la base interne à l'entreprise
Base documentaire associée : Codes sources de programmes écrits dans différents langages de programmation dans l'entreprise
Mode d'intégration : Développement en interne
Mode d'hébergement : L'outil de vectorisation et moteur de recherche vectoriel était en local (on-premise), l'appel au LLM est sur le *cloud*
La vectorisation et la recherche vectorielle se font avec des données non-chiffrées, tandis que l'appel au LLM s'effectue avec des données chiffrées. Cela permet de garantir que seules des données chiffrées quittent l'enceinte de l'entreprise pour aller dans le cloud.

2. Assistants de gestion de la connaissance

Cas d'usage : Assistants pouvant être déployé à la demande des unités de l'entreprise pour de multiples cas d'usage. Par exemple :

- Assistant pour le département scientifique (recherche et synthèse d'informations issues d'articles, de thèses et de descriptions de procédés industriels)
- Assistant RH (appariement entre offres d'emplois et profils de compétences)
- Assistant pour le département marketing (analyses concurrentielles)

Base documentaire associée : une base par unité utilisatrice (ex : documentation scientifique, documents RH, documents financiers d'une entité…)
Mode d'intégration : Développement en interne
Mode d'hébergement : *Cloud*
Spécificités techniques : Système avancé d'évaluation avec une évaluation humaine spécifique à chaque cas d'usage.

#DGEntreprises

→ www.entreprises.gouv.fr

@DGEntreprises

Conception : SIRCOM/ Studio graphique – Crédit photo : monsitj