keyrus
make data matter

Des data RAGs aux richesses de l'IA

Guide opératoire pour suivre votre checklist

www.keyrus.com

Des data RAGs aux richesses de l'IA

Guide opératoire pour suivre votre checklist

La génération augmentée par récupération, plus connue sous l'acronyme RAG permet de produire des réponses contextualisées et adaptées aux besoins opérationnels des entreprises. Contrairement aux modèles d'IA classiques, souvent limités à des connaissances générales et figées, le RAG tire parti des bases de données internes pour fournir des informations précises, pertinentes et actualisées. Cette capacité en fait un levier stratégique majeur pour les organisations cherchant à optimiser leurs processus et à valoriser leurs données.

Les avantages du RAG sont multiples : gains de productivité, amélioration de la qualité des tâches réalisées et soutien à la formation et à l'acculturation numérique des équipes. Cependant, la mise en œuvre d'un système RAG ne s'improvise pas. Elle nécessite une approche structurée, impliquant des choix techniques et organisationnels adaptés aux spécificités de chaque entreprise.

Ce guide a été conçu pour accompagner les entreprises dans cette démarche. Il s'adresse aux décideurs, directeurs de projets, responsables de la transformation digitale et toutes les parties prenantes souhaitant exploiter pleinement le potentiel du RAG. À travers une checklist en 10 points clés et un plan de mise en œuvre détaillé, nous vous proposons un cadre opérationnel clair pour réussir l'intégration du RAG dans vos activités.

Nous commencerons par expliquer les principes fondamentaux du RAG et les bénéfices qu'il peut apporter. Ensuite, nous détaillerons chaque étape nécessaire pour transformer vos données en richesse : de l'identification des cas d'usage à la gestion des enjeux de sécurité et de conformité, en passant par le choix des technologies, la préparation des données et la formation des utilisateurs. Ce guide se veut à la fois méthodique et pratique, pour vous permettre de passer de la réflexion à l'action.

Images générées par **Midjourney** avec la consigne : *« Hyper-realistic portrait of a dark haired woman in a calm attitude like in a library, capturing the essence with light eyes, lit by overhead lighting , framed in a centered manner --chaos 10 --ar 3:2 --style raw --v 6 »*

Comprendre le RAG

La génération augmentée par récupération (*Retrieval-Augmented Generation* ou RAG) est une technologie qui combine l'intelligence artificielle générative et la récupération d'informations dans des bases de données internes. Elle se distingue des modèles d'intelligence artificielle traditionnels, qui s'appuient exclusivement sur des connaissances générales acquises lors de leur entraînement. Avec le RAG, les modèles génératifs, tels que les grands modèles de langage (LLM), consultent des informations actualisées et spécifiques à l'entreprise pour répondre de manière contextualisée aux besoins métiers.

Un système RAG repose sur deux modules principaux :

- **Le module de récupération d'informations** : il interroge les bases de données internes pour extraire les documents ou segments les plus pertinents en fonction d'une requête.
- **Le module de génération** : il utilise un modèle d'IA générative pour transformer ces informations en réponses en langage naturel, adaptées au contexte de la requête.

Cette combinaison permet d'assurer une meilleure précision des réponses, tout en réduisant les risques d'hallucination (réponses erronées ou incohérentes). Le RAG garantit également une traçabilité des sources utilisées, un atout majeur pour la fiabilité des systèmes.

Pourquoi utiliser le RAG ?

Le RAG répond à des besoins stratégiques pour les entreprises, notamment dans des environnements où les informations doivent être :

- **Précises et contextualisées** : en intégrant les connaissances internes, le RAG s'assure que les réponses sont adaptées aux spécificités de l'entreprise.
- **Actualisées** : contrairement aux modèles figés lors de leur entraînement, le RAG peut accéder à des données en temps réel.
- **Traçables** : les références des documents utilisés pour générer les réponses sont clairement identifiées, ce qui renforce la confiance des utilisateurs.

Les bénéfices du RAG se déclinent en trois axes principaux :

- **Gains de productivité** : en automatisant des tâches telles que la rédaction d'e-mails, de comptes rendus ou la recherche d'informations complexes, par exemple, le RAG permet de libérer du temps pour des missions à plus forte valeur ajoutée.
- **Amélioration de la qualité du travail** : grâce à une recherche exhaustive et rapide, les utilisateurs disposent d'une information complète et pertinente.
- **Soutien à la formation et à la montée en compétences** : le RAG peut servir de plateforme de partage de connaissances, favorisant l'apprentissage continu des équipes.

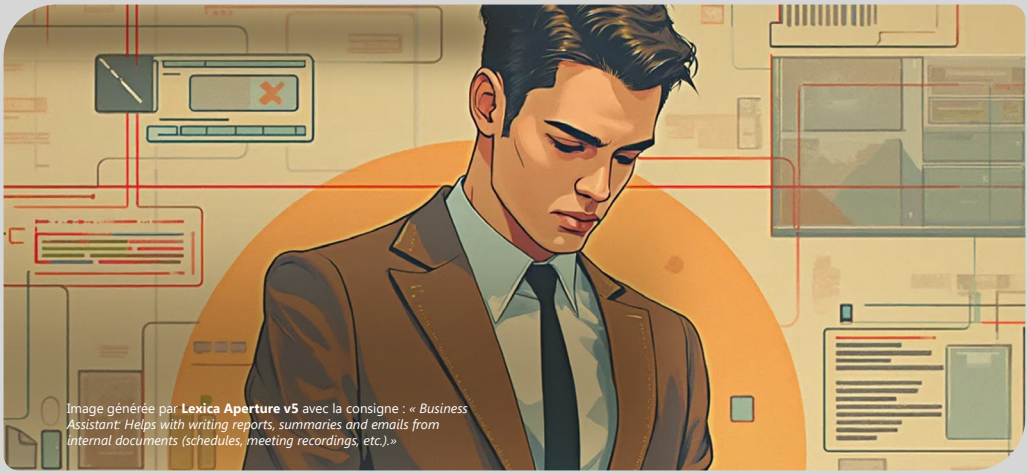


Image générée par **Lexica Aperture v5** avec la consigne : *« Business Assistant: Helps with writing reports, summaries and emails from internal documents (schedules, meeting recordings, etc.).»*

Cas d'usage typiques

Le RAG est particulièrement adapté aux tâches nécessitant une interaction avec des bases de données internes et des réponses en langage naturel. Voici quelques exemples concrets :

- **Assistant d'entreprise** : aide à la rédaction de comptes rendus, synthèses et e-mails à partir des documents internes (plannings, enregistrements de réunions, etc.)
- **Assistant juridique** : vérification de la conformité ou aide à la rédaction de contrats en s'appuyant sur des bases de données juridiques internes.
- **Assistant RH** : rédaction de documents RH, recherche dans les données internes sur les employés, appariement entre fiches de poste et CV.
- **Recherche dans la documentation technique** : extraction et synthèse des informations issues de manuels ou de documents spécialisés pour des bureaux d'études ou des équipes de maintenance.
- **Création de nouveaux produits** : analyse des historiques de conception et des échanges passés pour identifier des opportunités d'innovation.

Cette liste n'est évidemment pas exhaustive ! *Sky is the limit* disait-on il y a quelques années. *Foreign galaxies are the limit* serions-nous tentés de dire aujourd'hui !

En quoi le RAG est-il innovant ?

Le RAG se distingue par son approche hybride, combinant :

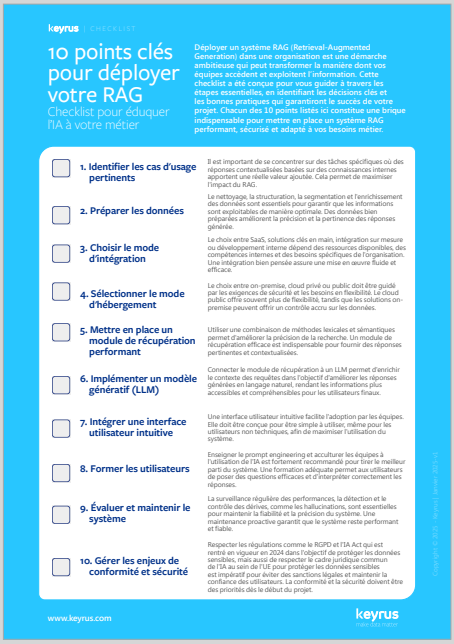
- **La puissance des modèles génératifs** pour la compréhension et la génération de langage naturel.
- **La pertinence des informations internes** grâce à des méthodes avancées de recherche et d'indexation des données.

Cette technologie est accessible à toutes les entreprises, quels que soient leur taille et leur secteur d'activité. Elle ne nécessite pas forcément d'expertise interne en intelligence artificielle, grâce à des solutions clés en main disponibles sur le marché. De plus, elle peut être adaptée pour répondre à des besoins très spécifiques, en s'appuyant sur des développements sur mesure.

Le RAG ne se contente pas de répondre à une demande : il augmente les capacités des utilisateurs en leur offrant des outils puissants pour exploiter leurs données de manière efficace et stratégique.

Guide de mise en œuvre opérationnelle de la checklist

Après avoir compris les principes fondamentaux et les bénéfices du RAG, il est temps de passer à l'action. Ce guide opérationnel se veut le complément de la checklist en ***10 points clés pour déployer votre RAG***. Il en constitue le prolongement pratique, en détaillant les étapes nécessaires pour transformer vos données en richesse grâce à cette technologie. Chaque étape de cette checklist est ici développée pour vous fournir les outils et les recommandations concrètes à chaque phase du projet.


keyrus | CHECKLIST

10 points clés pour déployer votre RAG

Checklist pour éduquer l'IA à votre métier

Déployer un système RAG (Retrieval-Augmented Generation) dans une organisation est une démarche ambitieuse qui peut transformer la manière dont vos équipes accèdent et exploitent l'information. Cette checklist a été conçue pour vous guider à travers les étapes essentielles, en identifiant les décisions clés et les bonnes pratiques qui garantiront le succès de votre projet. Chacun des 10 points listés ici constitue une brique indispensable pour mettre en place un système RAG performant, sécurisé et adapté à vos besoins métier.

1. Identifier les cas d'usage pertinents
Il est important de se concentrer sur des tâches spécifiques où des réponses contextualisées basées sur des connaissances internes apportent une réelle valeur ajoutée. Cela permet de maximiser l'impact du RAG.

2. Préparer les données
Le nettoyage, la structuration, la segmentation et l'enrichissement des données sont essentiels pour garantir que les informations sont exploitables de manière optimale. Des données bien préparées amélioreront la précision et la pertinence des réponses générée.

3. Choisir le mode d'intégration
Le choix entre SaaS, solutions clés en main, intégration sur mesure ou développement interne dépend des ressources disponibles, des compétences internes et des besoins spécifiques de l'organisation. Une intégration bien pensée assure une mise en œuvre fluide et efficace.

4. Sélectionner le mode d'hébergement
Le choix entre on-premise, cloud privé ou public doit être guidé par les exigences de sécurité et les besoins en flexibilité. Le cloud public offrir souvent plus de flexibilité, tandis que les solutions on-premise peuvent offrir un contrôle accru sur les données.

5. Mettre en place un module de récupération performant
Utiliser une combinaison de méthodes lexicales et sémantiques permet d'améliorer la précision de la recherche. Un module de récupération efficace est indispensable pour fournir des réponses pertinentes et contextualisées.

6. Implémenter un modèle génératif (LLM)
Connecter le module de récupération à un LLM permet d'enrichir le contexte des requêtes dans l'objectif d'améliorer les réponses générées en langage naturel, rendant les informations plus accessibles et compréhensibles pour les utilisateurs finaux.

7. Intégrer une interface utilisateur intuitive
Une interface utilisateur intuitive facilite l'adoption par les équipes. Elle doit être conçue pour être simple à utiliser, même pour les utilisateurs non techniques, afin de maximiser l'utilisation du système.

8. Former les utilisateurs
Enseigner le prompt engineering et acculturer les équipes à l'utilisation de l'IA est fortement recommandé pour tirer le meilleur parti du système. Une formation adéquate permet aux utilisateurs de poser des questions efficaces et d'interpréter correctement les réponses.

9. Évaluer et maintenir le système
La surveillance régulière des performances, la détection et le contrôle des dérives, comme les hallucinations, sont essentielles pour maintenir la fiabilité et la précision du système. Une maintenance proactive garantit que le système reste performant et fiable.

10. Gérer les enjeux de conformité et sécurité
Respecter les régulations comme le RGPD et l'IA Act qui est rentré en vigueur en 2024 dans l'objectif de protéger les données sensibles, mais aussi de respecter le cadre juridique commun de l'IA au sein de l'UE pour protéger les données sensibles est impératif pour éviter des sanctions légales et maintenir la confiance des utilisateurs. La conformité et la sécurité doivent être des priorités dès le début du projet.

www.keyrus.com

keyrus
make data matter

1: Identifier les cas d'usage pertinents

Analyser les besoins de l'entreprise

Avant de déployer un système RAG, il est essentiel de dresser un état des lieux précis des processus métiers existants pour identifier les opportunités d'amélioration :

Cartographie des processus métiers : listez les principales activités réalisées au sein de votre organisation et identifiez celles qui nécessitent un accès rapide à des informations internes. Les domaines tels que la rédaction de rapports, la gestion documentaire ou l'analyse technique sont souvent de bons points de départ.

- **Identification des tâches chronophages** : posez-vous des questions clés pour évaluer où le RAG peut avoir un impact significatif :
  - *« Quelles tâches consomment beaucoup de temps pour mes équipes ? »*
  - *« Quels processus nécessitent des recherches régulières dans des bases de données internes ? »*
  - *« Quelles sont les zones d'inefficacité liées à des informations mal organisées ou difficiles d'accès ? »*

Cette analyse vous permettra de cibler les processus où le RAG peut répondre à des besoins concrets et immédiats.

Prioriser les cas d'usage

Tous les cas d'usage identifiés ne se valent pas en termes de faisabilité ou de bénéfices attendus. Une priorisation est donc indispensable pour maximiser l'impact du RAG dès les premières phases de déploiement :

- **Évaluation de l'impact potentiel** :
  - Mesurez le gain de temps attendu sur les tâches identifiées.
  - Évaluez l'amélioration de la qualité des résultats obtenus grâce au RAG, notamment dans des activités nécessitant une recherche approfondie ou une grande précision.
  - Quantifiez la réduction des coûts opérationnels, par exemple en automatisant des tâches répétitives.

- **Analyse de la complexité de mise en œuvre** :
  - Estimez la facilité d'accès et de structuration des données nécessaires pour le cas d'usage.
  - Prenez en compte les compétences internes disponibles pour configurer et exploiter le système RAG.
- **Cibler les processus critiques** : privilégiez les cas d'usage ayant un fort impact stratégique ou opérationnel. Les tâches simples mais récurrentes, comme la recherche d'informations dans des bases documentaires ou la rédaction automatique d'e-mails, sont souvent des points de départ idéaux pour initier un projet RAG.
- **Hiérarchiser selon le ROI attendu** : mettez en avant les processus offrant un retour sur investissement rapide, afin de démontrer rapidement la valeur ajoutée du RAG à vos parties prenantes.

2 : Préparer les données

La préparation des données constitue une étape fondamentale pour garantir le bon fonctionnement et la performance de votre système RAG. Un nettoyage et une structuration minutieux permettent d'optimiser la précision des réponses générées, tout en assurant la conformité réglementaire et la sécurité des informations.

- **Élimination des doublons** : les fichiers redondants ou dupliqués représentent une source fréquente d'incohérences dans les bases de données. Les doublons peuvent fausser les résultats et alourdir inutilement la base documentaire, ce qui impacte la rapidité et la pertinence des recherches. Comment procéder ? Utilisez des outils d'analyse pour détecter et supprimer les doublons (ex. : algorithmes de déduplication basés sur des empreintes numériques ou des similitudes textuelles). Automatisez ce processus pour garantir une base toujours propre lors de l'ajout de nouveaux documents.
- **Segmentation des documents** : elle consiste à diviser vos documents en segments plus petits, appelés « chunks », qui seront exploités par le module de récupération. Pourquoi c'est important ? Un document entier peut être trop volumineux ou imprécis pour une recherche efficace. La segmentation permet de récupérer uniquement la partie pertinente de chaque document. Comment procéder ? Définissez des règles de segmentation basées sur la structure des documents (paragraphes, chapitres, titres). Assurez-vous que chaque chunk conserve le contexte nécessaire (par exemple, incluez des métadonnées comme le titre du document et sa date). Testez les tailles de chunks pour trouver un équilibre entre précision et vitesse de récupération (les tailles idéales varient souvent entre 300 et 500 mots).
- **Anonymisation ou pseudonymisation** : la protection des données sensibles est essentielle pour garantir la conformité avec les réglementations, comme le RGPD. Pourquoi c'est important ? L'utilisation de données personnelles non protégées peut entraîner des sanctions juridiques et nuire à la confiance des parties prenantes. Comment procéder ? De plusieurs manières ! **Par l'anonymisation** : Supprimez ou remplacez toute information permettant d'identifier une personne (ex. : noms, adresses, numéros de téléphone). Cette approche est irréversible.

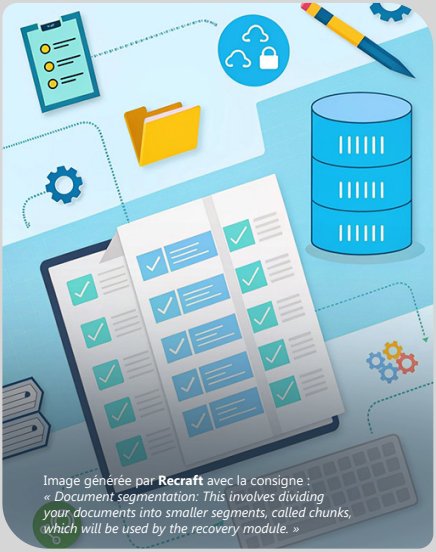


Image générée par **Recraft** avec la consigne : *« Document segmentation: This involves dividing your documents into smaller segments, called chunks, which will be used by the recovery module. »*

Par pseudonymisation : remplacez les données personnelles par des identifiants uniques réversibles (ex. : ID utilisateur). Cette méthode permet de rétablir les données d'origine si nécessaire, tout en réduisant les risques de confidentialité. Enfin, utilisez des outils automatisés pour détecter et masquer les informations sensibles dans vos documents.

Une base de données bien nettoyée et structurée est essentielle pour garantir le succès de votre système RAG. L'élimination des doublons, la segmentation des documents et la protection des données sensibles assurent des performances optimales tout en respectant les exigences réglementaires et éthiques.

- **Enrichissement et prétraitement des données** : ajoutez des métadonnées pertinentes à chaque segment de document (par exemple, type de document, date de création, sujet traité) pour faciliter la recherche. Si votre organisation dispose d'un vocabulaire ou de concepts spécifiques, créez un graphe de connaissances pour relier ces éléments et améliorer le contexte fourni au système. Enfin avec la vectorisation des données, vous convertissez chaque segment en un format mathématique (vecteurs) pour permettre une recherche rapide et efficace.

Après le nettoyage et la structuration, l'enrichissement et le prétraitement des données sont des étapes importantes pour optimiser leur exploitation par un système RAG. Ces processus permettent de rendre vos informations plus accessibles, compréhensibles et efficaces lors des recherches et des interactions avec le modèle génératif.

- **Ajout de métadonnées pertinentes**. L'ajout de métadonnées enrichit chaque segment de document en lui associant des informations contextuelles utiles, telles que le type de document, sa date de création ou le sujet traité. Ces métadonnées permettent au système RAG de mieux filtrer les résultats et de proposer des réponses plus précises et pertinentes. Pour ce faire, il est important de définir des métadonnées adaptées à vos besoins métier. Par exemple, un document peut être classé comme un contrat, un rapport ou une note interne, avec des catégories thématiques spécifiques. Ces métadonnées doivent être uniformes et suivre une nomenclature claire pour éviter les ambiguïtés. Leur extraction et leur ajout peuvent être automatisés à l'aide d'outils de gestion de contenu ou de scripts dédiés.
- **Création d'un graphe de connaissances**. Le graphe de connaissances est un outil puissant pour relier les concepts, entités et relations propres à votre organisation dans une structure interconnectée. Cette approche aide le système RAG à mieux comprendre le contexte des informations et à renforcer la pertinence des réponses générées. Par exemple, les produits, clients ou processus spécifiques à votre organisation peuvent être modélisés sous forme de nœuds, reliés par des relations explicites comme « Produit X appartient à la catégorie Y » ou « Client A a signé le contrat B ». La construction de ce graphe peut être réalisée à l'aide d'outils spécialisés comme Neo4j ou des bibliothèques Python dédiées. Ce graphe doit évoluer régulièrement pour refléter fidèlement les besoins métier et s'adapter aux nouvelles données.
- **Vectorisation des données**. La vectorisation consiste à convertir chaque segment de document en une représentation numérique, ou vecteur, qui peut être exploité par le moteur de recherche et le modèle génératif. Cette étape est essentielle pour permettre au système de comprendre le sens des segments et d'identifier ceux qui sont les plus pertinents par rapport à une requête utilisateur. Des modèles de vectorisation préentraînés, tels que Sentence-BERT ou OpenAI Embeddings, peuvent être utilisés pour cette tâche. Les vecteurs générés sont ensuite stockés dans une base de données vectorielle, comme FAISS ou Pinecone, pour permettre une recherche rapide et efficace. Il est essentiel de mettre à jour régulièrement cette base vectorielle pour inclure les nouveaux documents ou segments, garantissant ainsi la pertinence des recherches dans le temps.

3: Choisir le mode d'intégration

La troisième étape de la mise en œuvre d'un système RAG consiste à sélectionner le mode d'intégration le plus adapté aux besoins et aux capacités de votre organisation. Ce choix est déterminant pour assurer une mise en œuvre fluide et efficace, tout en respectant les contraintes opérationnelles. Deux grandes options s'offrent généralement aux entreprises : les solutions clés en main ou les solutions sur mesure.

- **Solutions clés en main** (SaaS ou licences logicielles). Les solutions clés en main, proposées sous forme de logiciels en tant que service (SaaS) ou de licences logicielles, se distinguent par leur simplicité et leur rapidité de déploiement. Ces options sont particulièrement adaptées aux entreprises qui ne disposent pas de compétences techniques internes ou qui souhaitent un retour sur investissement rapide. Elles offrent généralement des interfaces prêtes à l'emploi et une prise en charge complète des aspects techniques, réduisant ainsi le besoin de formation initiale. En revanche, leur capacité de personnalisation peut être limitée, ce qui pourrait constituer un frein pour les cas d'usage très spécifiques.
- **Solutions sur mesure** : elles permettent une personnalisation avancée pour répondre précisément aux besoins spécifiques de votre organisation. Ce type d'intégration peut être réalisé en interne, si votre entreprise dispose des ressources techniques nécessaires, ou par le biais d'un intégrateur externe spécialisé. Les solutions sur mesure sont particulièrement adaptées aux cas d'usage complexes ou aux secteurs d'activité nécessitant des adaptations métier précises. Toutefois, cette option implique souvent des délais plus longs et des coûts potentiellement plus élevés, notamment en raison des phases de conception, de développement et de tests.
- **Critères de décision** : pour choisir le mode d'intégration le plus pertinent, plusieurs critères doivent être pris en compte :
- **Ressources internes** : évaluez si votre organisation dispose des compétences techniques nécessaires pour gérer un développement sur mesure ou si une solution clés en main serait plus appropriée.

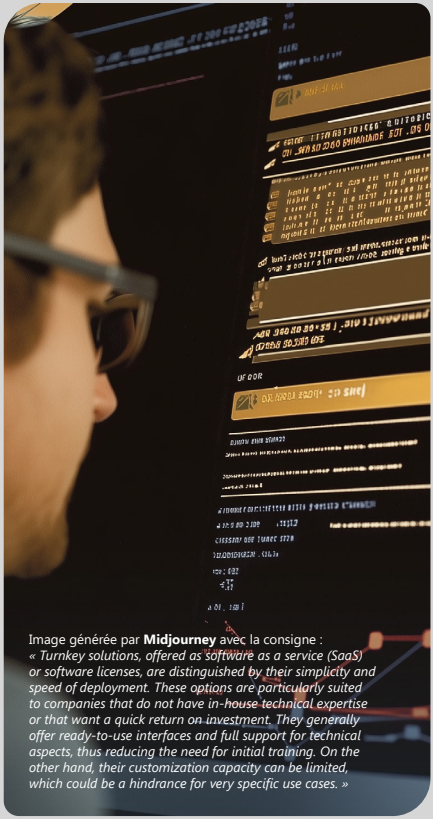


Image générée par **Midjourney** avec la consigne : *« Turnkey solutions, offered as software as a service (SaaS) or software licenses, are distinguished by their simplicity and speed of deployment. These options are particularly suited to companies that do not have in-house technical expertise or that want a quick return on investment. They generally offer ready-to-use interfaces and full support for technical aspects, thus reducing the need for initial training. On the other hand, their customization capacity can be limited, which could be a hindrance for very specific use cases. »*

- **Délais** : si vos objectifs nécessitent une mise en œuvre rapide, les solutions SaaS ou clés en main sont souvent plus adaptées.
- **Budget** : comparez les coûts initiaux et récurrents des deux options. Les solutions sur mesure, bien que plus flexibles, impliquent généralement un investissement financier plus important.
- **Complexité des cas d'usage** : si vos besoins sont standards, une solution clé en main suffira. En revanche, des exigences spécifiques ou des processus métier uniques justifient l'investissement dans une solution sur mesure.

Ainsi le choix du mode d'intégration est une décision stratégique qui doit être alignée avec vos ressources, vos contraintes et vos objectifs.

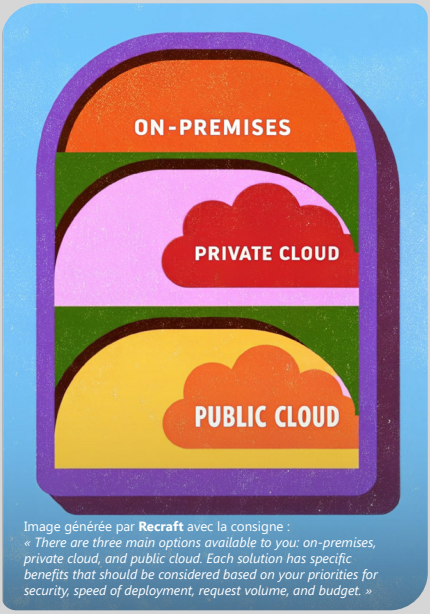


Image générée par **Recraft** avec la consigne :
« There are three main options available to you: on-premises, private cloud, and public cloud. Each solution has specific benefits that should be considered based on your priorities for security, speed of deployment, request volume, and budget. »

4: Évaluer et sélectionner le mode d'hébergement

Le choix du mode d'hébergement est une étape importante pour assurer la sécurité, la performance et l'évolutivité de votre système RAG. Selon vos besoins et contraintes, trois options principales s'offrent à vous : on-premise, cloud privé et cloud public. Chaque solution présente des avantages spécifiques qui doivent être examinés en fonction de vos priorités en termes de sécurité, de vitesse de déploiement, de volume de requêtes et de budget.

L'hébergement on-premise offre un contrôle total sur vos données et votre infrastructure. Il est particulièrement adapté aux organisations manipulant des informations hautement sensibles ou soumises à des réglementations strictes en matière de confidentialité. En optant pour cette solution, vous éliminez les risques liés à l'externalisation des données et vous bénéficiez d'une maîtrise complète des accès et des processus. Cependant, cette option implique des coûts initiaux importants pour l'acquisition et la maintenance de l'infrastructure, ainsi qu'un besoin accru en compétences internes pour la gestion technique.

Le cloud privé, quant à lui, constitue un bon compromis entre sécurité et flexibilité. Il permet de bénéficier d'une infrastructure dédiée à votre entreprise, garantissant une meilleure isolation des données et un contrôle renforcé par rapport au cloud public. Cette solution est idéale pour les organisations qui souhaitent tirer parti des avantages du cloud tout en préservant un niveau élevé de sécurité pour leurs données sensibles. Toutefois, le coût d'un cloud privé est généralement plus élevé que celui d'un cloud public, en raison de la personnalisation et de la gestion des ressources qui lui sont associées.

Enfin, **le cloud public** se distingue par sa rapidité de déploiement et ses coûts réduits. Cette option est particulièrement intéressante pour les entreprises ayant des contraintes budgétaires ou des besoins immédiats. Grâce à des certifications de sécurité, telles que le label SecNumCloud, le cloud public peut également répondre aux exigences des organisations en matière de protection des données. Cependant, comme les ressources sont partagées avec d'autres utilisateurs, cette solution est moins adaptée aux entreprises qui gèrent des informations sensibles ou critiques.

Critères clés : pour choisir le mode d'hébergement le plus adapté, plusieurs critères doivent être pris en compte. La sécurité des données est un facteur déterminant : les organisations manipulant des informations sensibles devraient privilégier l'on-premise ou le cloud privé. Si la rapidité de déploiement est un enjeu majeur, le cloud public s'impose souvent comme la meilleure option. Le volume de requêtes attendu doit également être évalué pour éviter toute saturation des infrastructures. Enfin, les contraintes budgétaires jouent un rôle essentiel dans la décision : le cloud public offre une solution économique à court terme, tandis que l'on-premise ou le cloud privé nécessitent un investissement initial plus conséquent.

En conclusion, le mode d'hébergement choisi doit refléter les priorités stratégiques de votre organisation. Tandis que l'on-premise garantit un contrôle total et une sécurité maximale, le cloud privé offre un équilibre entre personnalisation et flexibilité, et le cloud public privilégie la simplicité et les coûts réduits. Une évaluation approfondie de vos besoins vous permettra d'opter pour la solution la mieux adaptée à votre projet RAG.

Image générée par **Midjourney** avec la consigne :
« Document vectorization, performed during the data preparation step, converts segments into usable mathematical representations.. »

5: Mettre en place un module de récupération performant

Le module de récupération est l'un des piliers fondamentaux d'un système RAG. Il joue un rôle en identifiant les documents ou segments les plus pertinents en réponse à une requête. Pour garantir son efficacité, il est essentiel de choisir et de configurer les méthodes de récupération adaptées aux besoins de votre organisation. Cette étape inclut également la mise en place d'une base de données vectorielle robuste pour optimiser l'accès aux informations.

Choisir les méthodes de récupération

La récupération peut se faire à l'aide de plusieurs approches, chacune ayant ses spécificités et ses avantages.

1. **La recherche lexicale** repose sur l'utilisation de mots-clés pour retrouver les documents contenant les termes exacts de la requête. Cette méthode est simple à mettre en œuvre et adaptée aux cas d'usage où les termes de recherche sont précis et bien définis. Toutefois, elle montre rapidement ses limites face à la diversité linguistique, aux synonymes et aux formulations complexes. Par exemple, une recherche lexicale peut manquer des documents pertinents si les mots-clés employés ne correspondent pas exactement à ceux présents dans les données.
2. **La recherche sémantique**, en revanche, utilise des modèles avancés pour comprendre le sens et le contexte des requêtes et des documents. Elle permet de fournir des réponses plus précises et pertinentes, même lorsque les termes de recherche diffèrent légèrement. Cette approche est particulièrement utile dans les cas d'usage nécessitant une compréhension approfondie, comme la recherche dans des textes techniques ou des documents longs. Cependant, elle peut être plus exigeante en termes de ressources informatiques et de temps de configuration.
3. **L'approche hybride** combine les avantages des recherches lexicale et sémantique pour maximiser la qualité des résultats. En exploitant les forces des deux méthodes, cette approche garantit une récupération rapide et précise tout en tenant compte du contexte et de la diversité linguistique. Par exemple, une recherche hybride pourrait utiliser des mots-clés pour limiter la portée initiale de la recherche, puis appliquer des techniques sémantiques pour affiner les résultats.

Structure d'une base de données vectorielle robuste

Pour permettre une recherche efficace et rapide, il est essentiel de structurer une base de données vectorielle adaptée aux besoins du module de récupération. La vectorisation des documents, réalisée lors de l'étape de préparation des données, convertit les segments en représentations mathématiques exploitables. Ces vecteurs sont ensuite stockés dans une base optimisée pour les recherches, comme FAISS, Pinecone ou Milvus.

Une base vectorielle robuste facilite l'identification des documents les plus pertinents en réduisant le temps de recherche, même lorsque les bases de données sont volumineuses. Il est important de s'assurer que cette base est bien dimensionnée pour répondre aux besoins de votre organisation et qu'elle est régulièrement mise à jour pour inclure les nouvelles données.

6: Implémenter un modèle génératif (LLM)

Le modèle génératif (LLM, Large Language Model) constitue le cœur du système RAG, permettant de transformer les informations récupérées en réponses compréhensibles et contextualisées. Son implémentation est une étape clé pour assurer la pertinence et la qualité des interactions avec les utilisateurs. Afin de connecter efficacement le module de récupération au LLM, il faut suivre deux étapes fondamentales : la sélection d'un modèle adapté et la configuration optimale du système.

Sélectionner un LLM adapté aux besoins métier. Le choix du modèle génératif est central pour garantir que le système répond de manière adéquate aux spécificités de votre secteur et à vos cas d'usage. Plusieurs modèles sont disponibles sur le marché, chacun ayant ses points forts et ses limitations.

Pour sélectionner le LLM le plus pertinent :

- **Analysez vos besoins métier** : identifiez les types de questions et de tâches auxquelles le système devra répondre. Par exemple, un LLM capable de traiter des documents techniques ou juridiques devra avoir été entraîné sur des corpus spécialisés.
- É**valuez les options disponibles** : comparez les modèles en fonction de leur capacité à gérer le langage et les données propres à votre domaine. Des solutions comme GPT, Llama ou Claude proposent des fonctionnalités avancées, mais varient en termes de coût, de personnalisation et de compatibilité.
- **Testez les modèles** : effectuez des essais pour vérifier la pertinence des réponses générées par chaque modèle, en tenant compte de la complexité des cas d'usage identifiés.

Il est important de s'assurer que le modèle choisi est non seulement performant, mais également en mesure d'être intégré harmonieusement au module de récupération.

Configurer le système pour minimiser les erreurs et garantir la traçabilité. Une fois le modèle sélectionné, il est nécessaire de configurer le système pour optimiser son fonctionnement tout en limitant les risques d'erreurs, comme les hallucinations (réponses incorrectes ou incohérentes). Cette configuration inclut plusieurs éléments critiques :

- **Réduction des hallucinations** : ajustez les paramètres du LLM, tels que le degré de créativité (ou température), pour limiter les réponses spéculatives. Vous pouvez également imposer des contraintes sur l'utilisation des données récupérées, afin que le LLM se base uniquement sur des informations fiables.
- **Traçabilité des réponses** : configurez le système pour inclure des références explicites aux documents utilisés pour générer chaque réponse. Cela renforce la transparence et la confiance des utilisateurs envers le système.
- **Personnalisation des prompts** : optimisez les instructions envoyées au LLM (prompts) pour obtenir des réponses claires et pertinentes. Par exemple, guidez le modèle pour qu'il se concentre sur des informations spécifiques ou reformule les réponses dans un langage adapté à votre secteur.

Ces ajustements garantissent non seulement la qualité des réponses, mais aussi leur conformité avec les besoins métier et les exigences réglementaires.

7: Intégrer une interface utilisateur intuitive

L'adoption d'un système RAG repose en grande partie sur la facilité d'utilisation et l'ergonomie de son interface utilisateur. Une interface bien conçue joue un rôle clé pour garantir l'engagement des équipes et leur efficacité dans l'utilisation du système. En répondant aux besoins variés des utilisateurs et en proposant une expérience intuitive, vous maximisez l'impact du RAG au sein de votre organisation.

Pour garantir une adoption rapide, l'interface utilisateur doit être simple, claire et accessible, même pour les collaborateurs peu familiers avec les technologies d'intelligence artificielle. Une interface complexe ou peu intuitive risque de décourager les utilisateurs, limitant ainsi les bénéfices de la solution.

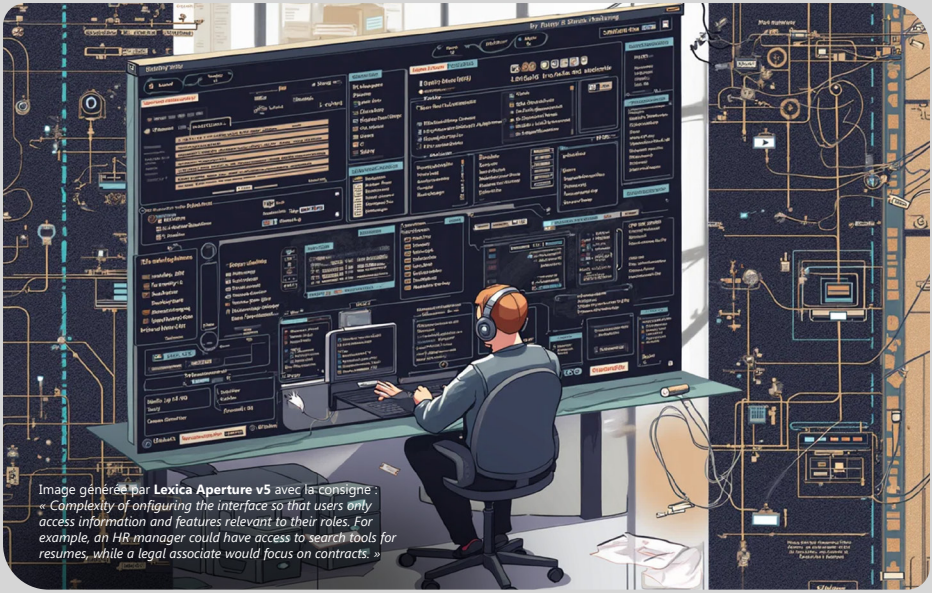


Image générée par **Lexica Aperture v5** avec la consigne : *« Complexity of onfiguring the interface so that users only access information and features relevant to their roles. For example, an HR manager could have access to search tools for resumes, while a legal associate would focus on contracts. »*

- **Simplicité d'utilisation** : l'interface doit privilégier des interactions claires et directes, avec des fonctionnalités faciles à comprendre et à manipuler. Par exemple, la saisie des requêtes doit être intuitive, et les résultats doivent être présentés de manière structurée, avec des options de tri et de filtrage.
- **Design épuré** : adoptez un design centré sur l'utilisateur, avec une navigation fluide et des visuels qui mettent en avant les informations essentielles. Les outils d'accessibilité, tels que des instructions guidées ou des tutoriels intégrés, peuvent également aider à faciliter la prise en main.
- **Compatibilité multiplateforme** : assurez-vous que l'interface fonctionne aussi bien sur un ordinateur que sur un appareil mobile ou une tablette, afin de permettre aux équipes de travailler dans différentes conditions.

Intégrez des options de personnalisation !
Les besoins des utilisateurs varient souvent en fonction de leurs rôles et responsabilités au sein de l'organisation. Une interface personnalisable permet d'adapter le système RAG aux préférences et aux exigences spécifiques de chaque groupe d'utilisateurs, renforçant ainsi leur engagement.

- **Rôles et permissions** : configurez l'interface pour que les utilisateurs accèdent uniquement aux informations et fonctionnalités pertinentes à leurs fonctions. Par exemple, un responsable RH pourrait avoir accès à des outils de recherche pour les CV, tandis qu'un collaborateur juridique se concentrerait sur les contrats.
- Tableaux de bord personnalisés : offrez aux utilisateurs la possibilité de personnaliser leur expérience en créant des tableaux de bord adaptés à leurs besoins spécifiques, incluant des indicateurs, des raccourcis ou des requêtes fréquentes.
- Langage et terminologie adaptés : si votre organisation utilise des termes ou des concepts spécifiques, veillez à les intégrer dans l'interface pour améliorer la compréhension et la pertinence des interactions.

En concevant une interface ergonomique, accessible et personnalisable, vous offrez à vos équipes un outil simple à utiliser et parfaitement adapté à leurs besoins. Cette approche ne se contente pas de faciliter l'usage du RAG : elle contribue également à renforcer la satisfaction des utilisateurs et à maximiser les bénéfices de la solution pour votre organisation.

8: Former les utilisateurs

La réussite de l'intégration d'un système RAG ne repose pas uniquement sur sa conception technique, mais aussi sur l'engagement et les compétences des équipes qui l'utilisent. Pour garantir une adoption efficace et durable, il est essentiel de mettre en place des actions d'acculturation et de formation. Ces initiatives permettent non seulement d'optimiser l'utilisation du système, mais aussi de renforcer la confiance des utilisateurs dans cette technologie.

Former au prompt engineering

Le prompt engineering est une compétence clé pour exploiter pleinement le potentiel d'un système RAG. Il s'agit d'apprendre aux utilisateurs à formuler des requêtes précises et efficaces pour maximiser la pertinence des réponses générées.

- **Pourquoi c'est important** : une question mal posée peut produire des réponses peu utiles ou incorrectes. En enseignant les meilleures pratiques pour structurer une requête, vous améliorez la qualité des interactions avec le système et augmentez la satisfaction des utilisateurs.
- **Comment procéder** : organisez des sessions de formation dédiées, adaptées aux niveaux de familiarité technologique des équipes. Ces formations peuvent inclure des exemples concrets, des exercices pratiques et des guides pour formuler des requêtes adaptées aux besoins métier. Vous pourriez, par exemple, former un juriste à poser des questions précises sur des clauses contractuelles ou un responsable RH à rechercher des profils spécifiques.

Démystifier l'IA et ses limites

Pour favoriser une adoption sereine du RAG, il est essentiel de communiquer de manière claire et transparente sur ce qu'il est capable de faire et, surtout, sur ses limites.

- **Insister sur le rôle d'assistance** : expliquez que le RAG est un outil conçu pour aider les équipes dans leurs tâches quotidiennes, mais qu'il ne remplace pas leur expertise. Cette distinction est importante pour rassurer les collaborateurs qui pourraient percevoir l'IA comme une menace pour leur emploi ou leurs compétences.
- **Clarifier les limites technologiques** : présentez les points faibles de l'IA, tels que les risques d'hallucination ou d'interprétation incorrecte des données, pour que les utilisateurs sachent comment détecter et gérer ces situations. Par exemple, montrez comment vérifier les références fournies par le RAG ou comment reformuler une requête si la réponse semble inexacte.
- **Encourager une posture critique** : invitez les utilisateurs à ne pas accepter les réponses générées comme des vérités absolues, mais à les utiliser comme un point de départ pour leurs analyses ou décisions.

Investir dans l'acculturation et la montée en compétences des équipes est un levier essentiel pour garantir le succès de votre projet RAG. En formant vos collaborateurs au prompt engineering et en les accompagnant pour mieux comprendre les capacités et les limites de l'IA, vous créez un environnement où l'outil est perçu comme un allié puissant.

9: Évaluer et maintenir le système

L'évaluation régulière des performances du système est essentielle pour détecter les problèmes et ajuster les paramètres. Voici les étapes clés :

- **Surveillance des indicateurs de performance** :
  - **Pertinence des réponses** : mesurez si les réponses générées répondent avec précision aux besoins des utilisateurs. Un faible taux de pertinence indique qu'un ajustement du module de récupération ou du modèle génératif est nécessaire.
  - **Taux d'hallucination** : identifiez les réponses incorrectes ou incohérentes produites par le modèle pour réduire les risques d'erreurs.
  - **Temps de réponse** : assurez-vous que les délais entre la requête utilisateur et la réponse générée sont optimaux, même en cas de forte sollicitation.

- **Recueil des retours utilisateurs** : impliquez les utilisateurs finaux dans le processus d'évaluation, à travers des enquêtes, des commentaires ou des sessions de test. Leur perception est un indicateur clé pour détecter des problèmes opérationnels ou des attentes non satisfaites.
- **Audit régulier des résultats** : comparez les réponses générées avec des données de référence pour identifier les écarts ou incohérences.

Et pour maintenir un système RAG performant face aux évolutions des données, des besoins métiers et des technologies, **la maintenance régulière** est indispensable :

- **Mise à jour des modèles** :
  - Réalisez des mises à jour périodiques du modèle génératif (LLM) pour intégrer les améliorations technologiques et corriger les bugs identifiés.
  - Adaptez le module de récupération si des ajustements sont nécessaires pour gérer des données nouvelles ou des formats variés.
- **Intégration des nouvelles données** :
  - Automatisez le processus d'ajout des nouvelles données dans la base documentaire, avec un passage systématique par le module de prétraitement (nettoyage, segmentation, vectorisation).
  - Assurez une vérification régulière de la qualité des données ajoutées pour éviter les incohérences ou les biais.
- **Correction des dérives** :
  - Identifiez les baisses de performance ou les anomalies (comme une augmentation du taux d'hallucination ou des erreurs fréquentes sur des cas spécifiques).
  - Adaptez les paramètres du système ou requalifiez les données problématiques pour corriger ces dérives.
- **Surveillance proactive des performances** :
  - Mettez en place des outils de monitoring en temps réel pour détecter rapidement toute dégradation du système.
  - Utilisez des tableaux de bord pour suivre les indicateurs clés et alerter les équipes en cas d'anomalie.

Images générées par **Midjourney** avec la consigne :
« Involving end users in the evaluation process, through surveys, feedback or testing sessions. --chaos 10 --ar 1:2 --style raw --weird 300 »

10 : Gérer les enjeux de conformité et sécurité

Et comment ne pas évoquer **la protection des données sensibles** ! Pour protéger vos informations critiques et prévenir les risques de fuite ou de cyberattaques, il est impératif de mettre en place des mesures adaptées :

- **Hébergement sécurisé** : privilégiez des infrastructures on-premise pour un contrôle total sur vos données, ou optez pour des services de cloud certifiés, tels que ceux labellisés SecNumCloud. Ces solutions offrent des garanties élevées en matière de sécurité et de conformité.
- **Cryptage des données** : assurez le chiffrement des données en transit et au repos. Cela protège les informations contre tout accès non autorisé, que ce soit au sein de l'entreprise ou lors d'échanges avec des tiers.

Le respect des normes et réglementations en vigueur est également un sujet incontournable pour éviter tout risque juridique et assurer la pérennité de votre projet.

- **Respect des lois locales** : garantissez la conformité avec le RGPD et les exigences du Règlement IA européen (AI Act). Ces cadres légaux encadrent l'utilisation des données personnelles et imposent des obligations spécifiques pour les systèmes d'intelligence artificielle, notamment en matière de transparence et d'évaluation des risques.
- **Auditabilité des processus** : documentez rigoureusement les étapes de traitement et d'utilisation des données. Maintenez un historique clair des interactions avec le système RAG pour faciliter les audits et renforcer la traçabilité.

Enfin, **adopter une démarche éthique** est très significatif pour renforcer la confiance des utilisateurs et garantir une utilisation équitable du système. Parmi les enjeux et/ou défis, évoquons notamment :

- **La réduction des biais** : identifiez et atténuez les biais dans les données et les modèles pour éviter les réponses discriminatoires ou injustes. Une analyse approfondie des sources et un contrôle continu des résultats sont nécessaires pour garantir une équité maximale.
- **L'explicabilité des décisions** : assurez-vous que le système peut fournir des explications claires sur les réponses générées, notamment en identifiant les sources et en détaillant les étapes du processus. Cette transparence est essentielle pour les utilisateurs finaux et pour répondre aux exigences réglementaires.

En conclusion, intégrer des pratiques robustes de sécurité et de conformité tout en adoptant une approche éthique est un gage de succès pour votre projet RAG. Ces mesures ne se contentent pas de protéger vos données : elles renforcent la fiabilité, la transparence et l'acceptabilité du système, autant auprès des utilisateurs que des autorités réglementaires.

Ainsi, la génération augmentée par récupération (RAG) représente une avancée majeure dans l'utilisation de l'intelligence artificielle en entreprise. En permettant de connecter des modèles génératifs à vos données internes, cette technologie offre une opportunité unique de valoriser vos informations, d'automatiser des tâches complexes et d'améliorer la qualité des décisions. Toutefois, pour exploiter pleinement son potentiel, une approche méthodique et structurée est indispensable.

Article co-écrit par Keyrus, ChatGPT4o et Mistral, Claude, Copilot, Perplexity et Gemini

Un accompagnement sur mesure

Ce guide vous a présenté une feuille de route opérationnelle en 10 étapes, couvrant l'ensemble du cycle de mise en œuvre du RAG, de l'identification des cas d'usage à la gestion des enjeux de conformité et de sécurité. En suivant cette checklist, vous pouvez :

- maximiser la pertinence et la précision des réponses générées,
- améliorer la productivité et l'efficacité de vos équipes,
- instaurer un cadre sécurisé et éthique pour l'utilisation de vos données.

Au-delà des bénéfices techniques et opérationnels, le déploiement du RAG marque un véritable tournant dans votre transformation numérique. Il s'agit d'un levier stratégique qui, bien intégré, peut renforcer votre compétitivité et accompagner la montée en compétences de vos équipes.

Chez Keyrus, nous comprenons les défis liés à l'intégration d'une nouvelle technologie comme le RAG. C'est pourquoi nous vous proposons un accompagnement sur mesure, adapté à vos besoins spécifiques et à vos objectifs. Que vous débutiez ou que vous souhaitiez perfectionner votre approche, nos experts sont là pour vous aider à tirer le meilleur parti de vos données grâce à des méthodologies véritablement éprouvées et des solutions innovantes.

Prenez dès aujourd'hui les premières mesures pour déployer le RAG dans votre organisation. Ensemble, transformons vos données en richesse et faisons de l'intelligence artificielle un atout clé pour votre entreprise.

Contactez-nous pour en savoir plus sur nos services ou pour bénéficier d'une expertise dédiée à votre projet.

Images générées par **Midjourney** avec la consigne :
« Data and AI business consultant --chaos 10 --ar 1:2 --style raw --weird 300»

Annexes

5 exemples de déploiement dans différents secteurs

Le RAG est une technologie adaptable à de nombreux secteurs d'activité, offrant des solutions concrètes pour des problématiques variées. Voici quelques exemples de déploiements réussis dans des domaines clés, illustrant comment cette technologie peut transformer les processus métiers et apporter une réelle valeur ajoutée.

1. Secteur juridique : assistance à la rédaction et à la vérification de contrats

Dans le domaine juridique, le RAG est particulièrement utile pour la gestion et l'analyse de documents complexes.

- **Cas d'usage** : un cabinet d'avocats peut s'appuyer sur un système RAG pour vérifier la conformité de contrats avec les réglementations en vigueur et rédiger des clauses adaptées à des besoins spécifiques.
- **Fonctionnalité principale** : le module de récupération extrait des segments pertinents de bases de données juridiques internes et des textes législatifs, tandis que le modèle génératif reformule les informations pour produire des clauses ou des synthèses compréhensibles.
- **Bénéfices** :
  - Réduction du temps consacré à la recherche documentaire.
  - Diminution des erreurs grâce à une traçabilité des informations utilisées.
  - Augmentation de la productivité des avocats et des juristes.

Image générée par **Midjourney** avec la consigne :
« Legal sector: assistance in drafting and verifying contracts. »

2. Ressources humaines : appariement des candidatures et gestion des processus de recrutement

Le RAG peut jouer un rôle déterminant dans l'optimisation des processus RH, notamment en matière de recrutement.

- **Cas d'usage** : une entreprise peut recourir au RAG pour appairer automatiquement des candidatures aux offres d'emploi internes, en tenant compte des compétences, de l'expérience et des préférences des candidats.
- **Fonctionnalité principale** : le système analyse les CV et les fiches de poste, identifie les correspondances pertinentes et génère des synthèses pour les responsables RH.
- **Bénéfices** :
  - Gain de temps dans le tri et la présélection des candidatures.
  - Meilleure adéquation entre les profils sélectionnés et les besoins de l'entreprise.
  - Amélioration de l'expérience candidat grâce à un processus plus rapide et précis.

Image générée par **Recraft** avec la consigne : *« Health sector: support for research and medical documentation. »*

3. Secteur commercial : personnalisation des offres et optimisation des ventes

Dans le secteur commercial, le RAG est un outil puissant pour améliorer l'expérience client et maximiser les opportunités de vente.

- **Cas d'usage** : une entreprise de retail peut utiliser le RAG pour générer des recommandations de produits personnalisées en fonction de l'historique d'achat et des préférences des clients.
- **Fonctionnalité principale** : le système récupère des données issues des interactions clients (commandes, e-mails, avis) et génère des propositions adaptées, comme des offres promotionnelles ou des conseils d'achat.
- **Bénéfices** :
  - Augmentation du taux de conversion grâce à des recommandations pertinentes.
  - Fidélisation des clients par une approche personnalisée.
  - Meilleure gestion des stocks en anticipant les besoins des consommateurs.

4. Secteur technique et industriel : recherche documentaire pour la maintenance

Dans les environnements industriels, le RAG peut simplifier l'accès aux informations techniques.

- **Cas d'usage** : un service de maintenance utilise le RAG pour rechercher des instructions spécifiques dans des manuels techniques volumineux lors de la réparation d'équipements.
- **Fonctionnalité principale** : le module de récupération extrait des segments précis des manuels ou des historiques de maintenance, tandis que le modèle génératif synthétise les informations pour guider les techniciens.
- **Bénéfices** :
  - Réduction des temps d'arrêt des machines grâce à un accès rapide aux informations critiques.
  - Amélioration de la précision des interventions techniques.
  - Optimisation de la formation des techniciens par une assistance en temps réel.

5. Secteur de la santé : soutien à la recherche et à la documentation médicale

Dans le domaine de la santé, le RAG peut accélérer la recherche et la gestion des connaissances médicales.

- **Cas d'usage** : un hôpital peut utiliser le RAG pour analyser des publications médicales et aider les cliniciens à prendre des décisions fondées sur les dernières avancées scientifiques.
- **Fonctionnalité principale** : le système récupère les articles pertinents dans des bases de données scientifiques et génère des résumés clairs pour les praticiens.
- **Bénéfices** :
  - Amélioration de la qualité des soins grâce à un accès rapide aux données actualisées.
  - Réduction du temps consacré à la recherche documentaire.
  - Meilleure prise de décision clinique basée sur des preuves.

Glossaire technique du RAG (*Retrieval-Augmented Generation*)

A

Augmentation : processus d'enrichissement des prompts avec des informations contextuelles pertinentes avant leur envoi au LLM.

API rate limiting : limitation du nombre de requêtes pouvant être effectuées vers une API (comme OpenAI) dans un intervalle de temps donné.

B

Batch processing : traitement des documents par lots pour optimiser les performances lors de l'ingestion de données.

BM25 (Best Match 25) : algorithme de ranking utilisé pour la recherche documentaire, alternative aux méthodes basées sur les embeddings.

C

Chunk : fragment de texte obtenu après découpage d'un document source, optimisé pour la recherche et la récupération.

Chunking : processus de découpage des documents en morceaux plus petits et gérables.

Context window : nombre maximum de tokens qu'un modèle peut traiter en une fois, incluant le prompt et la réponse.

Cosine similarity : mesure de similarité entre deux vecteurs, couramment utilisée pour comparer des embeddings.

D

Dense retrieval : méthode de recherche utilisant des représentations vectorielles denses (embeddings).

Document store : base de données optimisée pour stocker et récupérer des documents et leurs métadonnées.

E

Embedding : représentation vectorielle numérique d'un texte, capturant son sens sémantique.

Embedding model : modèle spécialisé dans la génération d'embeddings (ex: text-embedding-3-small d'OpenAI).

F

Fine-tuning : adaptation d'un modèle pré-entraîné à un domaine ou une tâche spécifique.

H

Hybrid Search : combinaison de plusieurs méthodes de recherche (ex: recherche sémantique + lexicale).

I

Index : structure de données optimisée pour la recherche rapide de documents ou de chunks.

Information retrieval : domaine concernant la recherche et l'extraction d'informations pertinentes depuis une base documentaire.

K

k-NN (k-Nearest Neighbors) : algorithme utilisé pour trouver les k documents les plus similaires à une requête.

L

LLM (Large Language Model) : modèle de langage large utilisé pour la génération de texte.

Latency : temps de réponse du système, nécessaire pour les applications en temps réel.

M

Metadata : informations supplémentaires associées aux documents (date, auteur, source, etc.)

N

Neural search : recherche basée sur des réseaux de neurones, typiquement via des embeddings.

P

Prompt engineering : conception et optimisation des prompts pour obtenir les meilleures réponses du LLM.

Prompt template : structure prédéfinie pour formater les prompts de manière cohérente.

Q

Query expansion : technique d'enrichissement des requêtes pour améliorer la pertinence des résultats.

R

RAG (Retrieval-Augmented Generation) : technique combinant la recherche documentaire et la génération de texte.

Reranking : processus de réordonnancement des résultats de recherche pour améliorer leur pertinence.

Retrieval : phase de recherche et récupération des documents pertinents.

S

Semantic search : recherche basée sur le sens plutôt que sur les mots exacts.

Sparse retrieval : méthode de recherche utilisant des représentations creuses (comme TF-IDF).

Image générée par **Recraft** avec la consigne *« Prompt engineering. »*

T

Text preprocessing : nettoyage et normalisation du texte avant son traitement.

Tokens : unités de texte utilisées par les modèles de langage (mots, sous-mots ou caractères).

Top-k : les k meilleurs résultats retournés par la recherche.

V

Vector database : base de données optimisée pour stocker et rechercher des embeddings.

Vector search : recherche basée sur la similarité entre vecteurs d'embeddings.

Concepts avancés

Cross-Encoder : modèle qui évalue la pertinence en prenant en compte simultanément la requête et le document.

Dense Passage Retriever (DPR) : architecture spécialisée dans la recherche de passages pertinents.

Maximum Inner Product Search (MIPS) : technique d'optimisation pour la recherche de vecteurs similaires.

Reciprocal Rank Fusion (RRF) : méthode de fusion des résultats de différents systèmes de recherche.

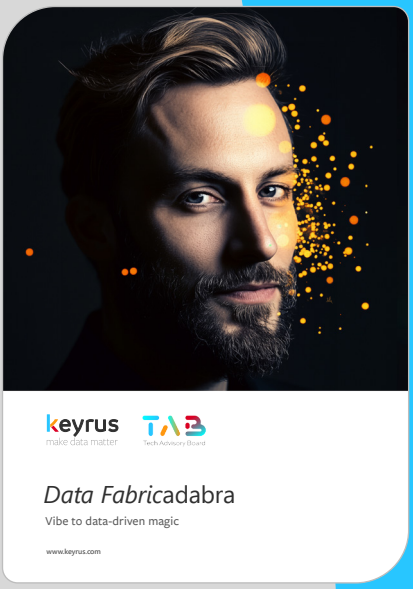
Vous avez trouvé cette lecture utile ?

Vous aimerez sûrement aussi :

Data Fabricadabra

Vibe to data-driven magic

Data Fabricadabra explore le concept de la Data Fabric, une architecture moderne qui transforme la gestion des données en un processus fluide et intégré. Cette solution permet de connecter des flux de données disparates et des environnements cloud, supprimant les silos de données et améliorant la prise de décision grâce à une vue unifiée des informations. En s'appuyant sur des technologies d'intelligence artificielle et d'automatisation, la Data Fabric offre une gestion des données en temps réel, simplifie la gouvernance et favorise une meilleure collaboration au sein des entreprises, tout en répondant aux défis de la transformation numérique et de l'augmentation des volumes de données.



Quelles sont les 3 principales idées ?

1. **Intégration et unification des données** : la Data Fabric connecte des sources de données variées pour offrir une vue unifiée en temps réel, facilitant la prise de décision.

2. **Automatisation et gouvernance renforcée** : elle utilise l'IA pour automatiser la gestion des données, garantissant leur qualité, sécurité et conformité.

3. **Flexibilité et évolutivité** : adaptable aux environnements hybrides et multi-cloud, la Data Fabric s'ajuste aux besoins croissants des entreprises en matière de gestion de données.

SCAN ME

keyrus

make data matter

Du conseil en management à 'intégration des technologies digitales, Keyrus met depuis 28 ans les données au cœur de chaque transformation pour accompagner ses clients dans l'amélioration continue et durable de leur performance au travers de 5 domaines d'expertise :

IA & Automatisation *: accompagner les entreprises et organisations publiques dans l'optimisation de leurs processus, l'augmentation de leur productivité et de leur performance opérationnelle pour leur permettre ainsi de se concentrer sur des activités à forte valeur ajoutée.*

Expérience digitale *: aider les entreprises à imaginer et à créer des expériences digitales multicanales inspirantes et engageantes pour atteindre leurs objectifs commerciaux.*

Data & Analytics *: permettre aux organisations de développer et de déployer les capacités nécessaires pour donner du sens et de la valeur aux données.*

Cloud & Sécurité *: offrir des solutions Cloud robustes, flexibles et sécurisées, garantissant la confidentialité et l'intégrité des données dans un environnement en pleine transformation.*

Transformation & Innovation *: aider les organisations à accélérer leur transformation métier et digitale et renforcer leur agilité, résilience et compétitivité dans un contexte en perpétuelle évolution.*

Avec une présence dans 28 pays et comptant plus de 3 300 experts, Keyrus est l'acteur incontournable et inspirant dans les domaines du conseil en management, de la data, du digital et bien sûr de l'IA en France et à l'international.

Pour en savoir plus : **www.keyrus.fr**

#HumanizingTheFuture

Jean-Philippe CLAIR
Directeur Marketing, Communication & Expérience client
jean-philippe.clair@keyrus.com