Réalisation d’une application de type RAG

**      **

Certification : Développeur d’application en Intelligence Artificiel

Marseille - Nice

LOMBARDI Joachim

Table des matières

[1. Réaliser la collecte, le stockage et la mise à disposition des données d’un projet en intelligence artificielle 3](#_Toc185683848)

[I. Programmer la collecte de données depuis plusieurs sources pour un projet en intelligence artificielle 3](#_Toc185683849)

# Réaliser la collecte, le stockage et la mise à disposition des données d’un projet en intelligence artificielle

## Programmer la collecte de données depuis plusieurs sources pour un projet en intelligence artificielle

### Identification des contraintes techniques propres aux sources de données

Tout média numérique est susceptible d’être utilisée comme source de donnée, les documents bien évidemment, mais aussi les sites internet, les vidéos, les bandes sonores, les images…

Afin de pouvoir en extraire des données, il faut d’abord connaître les règles de confidentialité.

#### Règles de confidentialité

Ces règles servent à protéger les données personnelles comme :

* Le nom, le prénom, l’adresse, le numéro de téléphone.
* Les informations d'identification (numéro de carte d'identité, numéro de sécurité sociale, etc.).
* Les données numériques : adresses IP, cookies, géolocalisation.
* Les données sensibles : informations de santé, opinions politiques, orientation sexuelle, croyances religieuses, etc.

Selon la CNIL, les 7 grands principes des règles de protection des données personnelles sont les suivants :

**Principe 1** : Licéité, limitation des finalités, minimisation des données

Licéité : Les données personnelles ne doivent être collectées que pour des finalités légitimes, explicites et déterminées.

Limitation des finalités : Les données collectées ne peuvent être utilisées que pour les finalités pour lesquelles elles ont été collectées.

Minimisation des données : La quantité de données collectées doit être réduite au minimum nécessaire pour atteindre les finalités déterminées.

En d'autres termes, les entreprises doivent être transparentes sur la raison pour laquelle elles collectent des données et ne collecter que les informations dont elles ont réellement besoin. Les mentions légales sur votre site internet sont là pour les informer.

**Principe 2** : Exactitude

Les données personnelles doivent être exactes et, si nécessaire, mises à jour. Des mesures doivent être prises pour que les données inexactes soient rectifiées ou effacées.

Cela signifie que les entreprises doivent prendre des mesures pour s'assurer que les données personnelles qu'elles détiennent sont exactes et à jour.

**Principe 3** : Limitation du traitement

Les données personnelles ne doivent être conservées que pendant une durée proportionnelle aux finalités pour lesquelles elles sont traitées.

En d'autres termes, les entreprises doivent avoir une politique claire de conservation des données et supprimer les données lorsqu'elles ne sont plus nécessaires.

**Principe 4** : Intégrité et confidentialité

Les données personnelles doivent être traitées de manière à garantir leur sécurité, y compris la protection contre le traitement non autorisé ou illicite et contre la perte, la destruction ou les dommages accidentels, par des mesures techniques ou organisationnelles appropriées.

Cela signifie que les entreprises doivent mettre en place des mesures de sécurité adéquates pour protéger les données personnelles contre les accès non autorisés, les utilisations abusives et les pertes accidentelles.

**Principe 5** : Responsabilisation

Le responsable du traitement est responsable du respect des principes susmentionnés et doit être en mesure de démontrer sa conformité.

Cela signifie que les entreprises doivent désigner un responsable de la protection des données (DPO) qui sera responsable de la mise en œuvre et du respect du RGPD.

**Principe 6** : Respect des droits des personnes

Les personnes ont le droit d'accéder à leurs données personnelles, de les rectifier, de les effacer, de limiter leur traitement, de s'opposer à leur traitement et de les transférer.

En d'autres termes, les entreprises doivent informer leurs clients de leurs droits en matière de protection des données et leur permettre d'exercer ces droits facilement.

**Principe 7** : Protection des données dès la conception et par défaut :

Le responsable du traitement doit mettre en œuvre des mesures techniques et organisationnelles appropriées pour protéger les données personnelles dès la conception du traitement et tout au long des produits et services.

Cela signifie que les entreprises doivent intégrer le traitement des données dans leur stratégie globale d'entreprise.

Ensuite, il existe également des contraintes lors de l’extraction des données :

#### Collecte des données

Comme indiqué dans l’introduction, il existe différents types de données outre les différents médias, un même type de source peux avoir plusieurs supports, par exemple ils peuvent être en format texte, JSON, CSV, base de données relationnelles… Ainsi, les méthodes d’extraction sont différentes et peuvent nécessiter du prétraitement comme pour les PDF afin d’avoir des données propres.

Ces méthodes d’extraction peuvent avoir des limites, comme lors du scraping (extraction de donnée depuis une page web) où le nombre de requêtes peut être limité, le téléchargement de fichier là aussi peut être limité ou également l’accès à des bases de données protégées par mot de passe. Il peut également y avoir des problèmes de collecte dû au réseau ou bien à la limitation imposée selon la région du monde où se trouve l’utilisateur ou bien les données.

Certaines données sont très volumineuses et imposent des contraintes au niveau de l’espace de stockage et la mémoire, nécessitant des ressources informatiques importantes et des optimisation de traitement.