

## Description de la Base de Données

La base de données utilisée pour notre système de recommandation d'emplois a été collectée à partir de plusieurs sources, notamment le site Glassdoor, ainsi que d'autres ressources externes. Une fois que les données ont été extraites, nous avons combiné ces informations et supprimé les doublons pour créer un ensemble de données plus vaste et représentatif. La version finale de la base de données contient **1450 offres d'emploi**.

### Origine des données

- **Source principale** : Les données ont été scrappées à partir du site **Glassdoor**, qui est une plateforme bien connue offrant des informations sur les entreprises, les salaires et les avis des employés.
- **Sources supplémentaires** : En complément, nous avons utilisé des données provenant d'autres sources comme ce lien Kaggle : <https://www.kaggle.com/datasets/rrkcoder/glassdoor-data-science-job-listings/data>, ainsi que d'autres ressources disponibles en ligne. Les informations de ces diverses sources ont été fusionnées pour garantir une couverture plus large du marché de l'emploi.

### Processus de prétraitement

- **Fusion des données** : Les différentes sources de données ont été combinées pour créer un ensemble de données unifié.
- **Suppression des doublons** : Afin de garantir l'intégrité des données et éviter la redondance, nous avons éliminé les doublons, ce qui a permis d'agrandir l'ensemble de données et d'augmenter la diversité des offres d'emploi.

### Caractéristiques initiales de la base de données

Le dataset initial contenait plusieurs caractéristiques qui fournissent des informations importantes sur chaque offre d'emploi. Voici les colonnes présentes dans les données avant toute modification :

1. **Job Title** : Le titre du poste ou du travail proposé (par exemple, "Data Scientist", "Software Engineer").
2. **Salary Estimate** : Estimation du salaire associé au poste, qui peut être donnée sous forme de salaire annuel ou d'une fourchette de salaires.
3. **Job Description** : Une description détaillée des tâches et responsabilités liées au poste.
4. **Rating** : Évaluation de l'entreprise sur la plateforme (note basée sur les avis des employés).
5. **Company Name** : Nom de l'entreprise proposant l'offre d'emploi.
6. **Location** : L'emplacement de l'emploi (ville, pays).
7. **Size** : La taille de l'entreprise, souvent exprimée en nombre d'employés.
8. **Founded** : L'année de création de l'entreprise.
9. **Type of Ownership** : Type de propriété de l'entreprise (privée, publique, etc.).

10. **Industry** : Secteur d'activité dans lequel l'entreprise opère (par exemple, technologie, finance, etc.).
11. **Sector** : Secteur spécifique dans lequel l'entreprise se classe (par exemple, services financiers, technologies de l'information, etc.).
12. **Revenue** : Les revenus annuels de l'entreprise, qui peuvent fournir une indication de sa taille et de sa réussite économique.

### Dimensions et Répartition des Données

- Le **dataset final** contient un total de **1450 offres d'emploi**, ce qui permet d'avoir une représentation suffisamment large des différentes industries, localisations géographiques et types d'entreprises.
- La base de données est bien équilibrée en termes de titres de poste, de salaires, et de localisation, ce qui la rend utile pour la génération de recommandations d'emploi personnalisées basées sur les profils des utilisateurs.

### Exploitation des données

Les données initiales sont extraites de manière à permettre une analyse détaillée des offres d'emploi. Cela inclut :

- **L'analyse des salaires** pour comprendre les tendances de rémunération en fonction des titres, secteurs ou localisations.
- **L'analyse des descriptions de poste** pour extraire des informations telles que les compétences clés, les expériences requises, et les qualifications.
- **L'évaluation des entreprises** à travers leur note, leur taille, et leur secteur, permettant de recommander des emplois en fonction des préférences de l'utilisateur pour certaines entreprises.

En résumé, cette base de données enrichie est un ensemble solide pour former un modèle de recommandation d'emploi efficace, qui pourra fournir des recommandations pertinentes et personnalisées aux utilisateurs sur la base de leur CV et de leurs préférences professionnelles.