

## Préparation des données pour clustering :

1. Jointure des bases de données **infront** pour construire une seule dataset « **Data\_Infront** » sur laquelle nous allons se baser :
  - « **Data\_Infront** » contient : **2637** Lignes, **270** colonnes.
  - « **Data\_Infront** » contient **2106** entreprises.
2. Suppression des colonnes et lignes dupliquées.
3. Gestion des valeurs nulles, colonne par colonne. (Certaines valeurs nulles seront supprimées, d'autres seront remplacées par des valeurs comme la moyenne, le médian, le max ...).
4. Gestion des outliers, colonne par colonne.
5. Normalisation des données.
6. Pour les colonnes contenant des textes comme « company description » :
  - Codage de ces colonnes
  - Construction d'un algorithme pour extraire de l'information de ces colonnes.

## Clustering (Classification non supervisée)