MU5RBC05

Génie logiciel et gestion de projet

Rapport du projet

Mise en œuvre d'une solution pour l'exploitation de données d'insertion professionnelle des étudiants de master

FERHAT Toufik
3872490
DAGMOUNE Asma
3703457

4 janvier 2022

Table des matières

1 Introduction

Les entreprises n'ont jamais eu autant de données à leur disposition, mais pour la plupart, elles ne savent pas quoi en faire, elles n'en utilisent qu'une infime partie. Le Big Data, c'est l'art d'utiliser efficacement l'énorme quantité de données dont dispose pratiquement chaque organisation. Le Big Data vous permet d'utiliser au mieux les données dont vous disposez, d'en extraire le maximum de valeur stratégique et de trouver plus facilement les réponses aux questions que vous vous posez.

L'objectif de ce projet est de pouvoir exploiter les données brutes sur l'insertion des étudiants de master dans le monde professionnel issues de data.gouv. Pour ce faire, il est nécessaire de faire une modélisation conceptuelle des données et de traduire cette modélisation conceptuelle en une modélisation logique afin de pouvoir stocker les données de manière efficace et de pouvoir les récupérer de manière tout aussi efficace.

Lorsque le processus de stockage est mis en place, nous devons être en mesure de comprendre l'utilité de ces données et de proposer des services et une visualisation.

Par manque de temps, nous n'avons pas implémenté une véritable interface graphique pour la visualisation, mais nous avons opté pour un Notebook Jupyter pour effectuer les requêtes et afficher les résultats.

2 Traitement des données brutes

2.1 Les données brutes

Les données traitées dans ce projet sont les résultats de plusieurs enquêtes qui ont été menées par le Ministère de l'Enseignement Supérieur sur l'insertion professionnelle des diplômés de master.
Ces résultats sont représentés au format CSV avec environ 15000 lignes et environ 30 colonnes qui représentent des statistiques et des attributs.
Ces données sont téléchargeables sous licence libre via ce lien.

2.2 Nettoyage des données

Dans la plupart des cas, les données brutes qui sont contenues dans des fichiers CSV ou d'autres types de fichiers structurés nécessitent une étape de pré-traitement pour être prêtes à être stockées dans une base de données prédéfinie.
Les données du projet n'étant pas une exception, on a dû aussi faire un nettoyage :

- Après une analyse manuelle du fichier CSV, nous avons remarqué que certaines cellules contenaient des valeurs telles que "ns" pour non significatif, "nd" pour non disponible et "." pour les salaires vides ou tout simplement des valeurs vides ; ces valeurs n'étant pas très utiles, nous avons choisi d'insérer ces données dans les tables sous forme de valeurs nulles.
- Nous avons également remarqué qu'il y avait une colonne intitulée "remarque" qui était dans la plupart des cas vide, mais qui donnait une indication sur la qualité et la fiabilité de l'enquête. Nous avons supprimé toutes les lignes qui comportaient une cellule "remarque" non nulle, car elle n'est non nulle que lorsque les données sont insuffisantes pour être exploitées.
- Nous avons également supprimé toutes les lignes où, selon notre modélisation, au moins une des clés primaires étaient nulles.

Tout le prétraitement des données a était réalisé en python à l'aide du framework pandas.

3 Modélisation et stockage des données

La modélisation des données permet de mieux structurer les données et de définir des contraintes d'intégrités qui facilitent le stockage et l'extraction des données.

3.1 Modélisation conceptuelle des données

La modélisation conceptuelle des données s'agit généralement d'une modélisation de premier niveau, dont les détails sont insuffisants pour construire une base de données réelle. Ce niveau décrit la structure de l'ensemble de la base de données pour un groupe d'utilisateurs. Il décrit les entités réelles et leurs relations.
Après réflexion et analyse du problème, on a pu concevoir la modélisation ci-dessous :

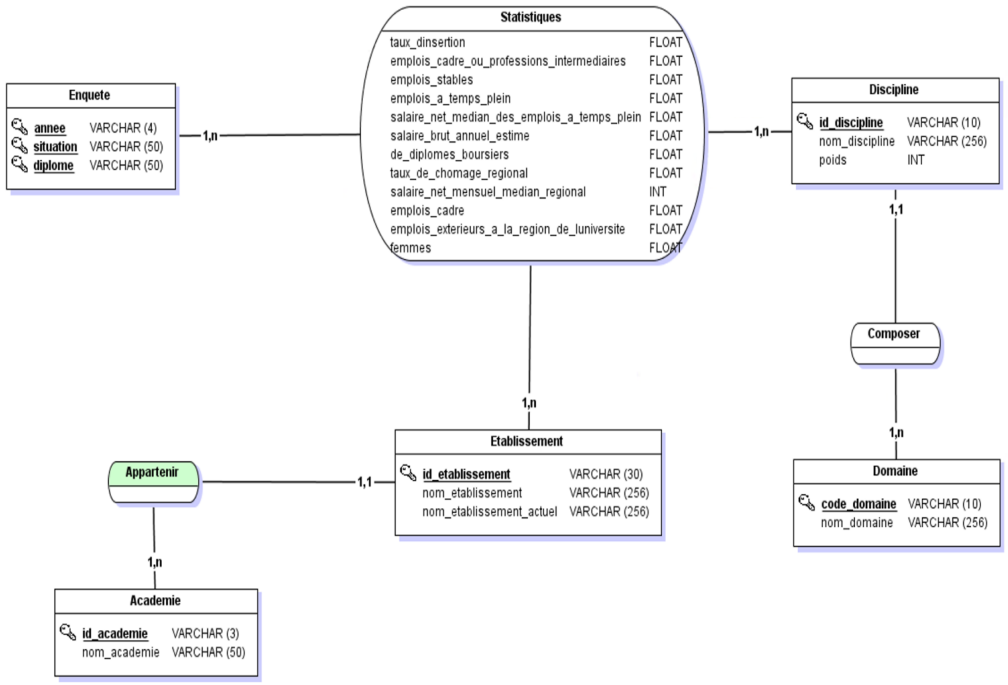

FIGURE 1 – Modélisation conceptuelle des données

Par rapport à notre compréhension du problème, nous avons identifié 5 entités :

- **l'enquête** qui est définie par l'année dont elle a était faite, le type de diplôme visé et la situation qui est de 18 mois ou 30 mois après le diplôme.
- **l'établissement** qui est l'université qui procure le diplôme, elle a comme identifiant le numéro d'établissement
- **l'académie** qui a validé le diplôme qui est identifié par le code de l'académie
- **le domaine** du diplôme qui a comme clé le code du domaine
- **la discipline** exacte du diplôme

Il existe bien sûr des relations entre la plupart des entités. Un établissement appartient nécessairement à une académie, il y a donc une relation d'appartenance. Une académie peut avoir plusieurs établissements, mais un établissement ne peut appartenir qu'à une seule académie, d'où la cardinalité dans le schéma.

Un domaine est composé de plusieurs disciplines, c'est plus ou moins la même relation qui existe entre l'académie et l'établissement, mais une discipline ne peut avoir qu'un seul domaine.

L'enquête et l'établissement sont liés par les statistiques réalisées, une enquête peut générer plusieurs statistiques d'un établissement particulier, de même, un établissement peut voir plusieurs statistiques réalisées par plusieurs enquêtes.

La discipline et l'enquête sont également associées à l'aide de statistiques et avec la même cardinalité.

3.2 Modélisation logique des données

La modélisation des données permet de traduire la modélisation conceptuelle en un format plus exploitable selon un système de gestion de base de données choisi.
Elle permet de spécifier exactement le type de données. Le passage du MCD au MLD nécessite le respect de certaines règles.
Comme notre MCD a été réalisé avec l'outil jMerise, le passage est automatique.
Le MLD que nous avons pu générer est disponible ci-dessous :

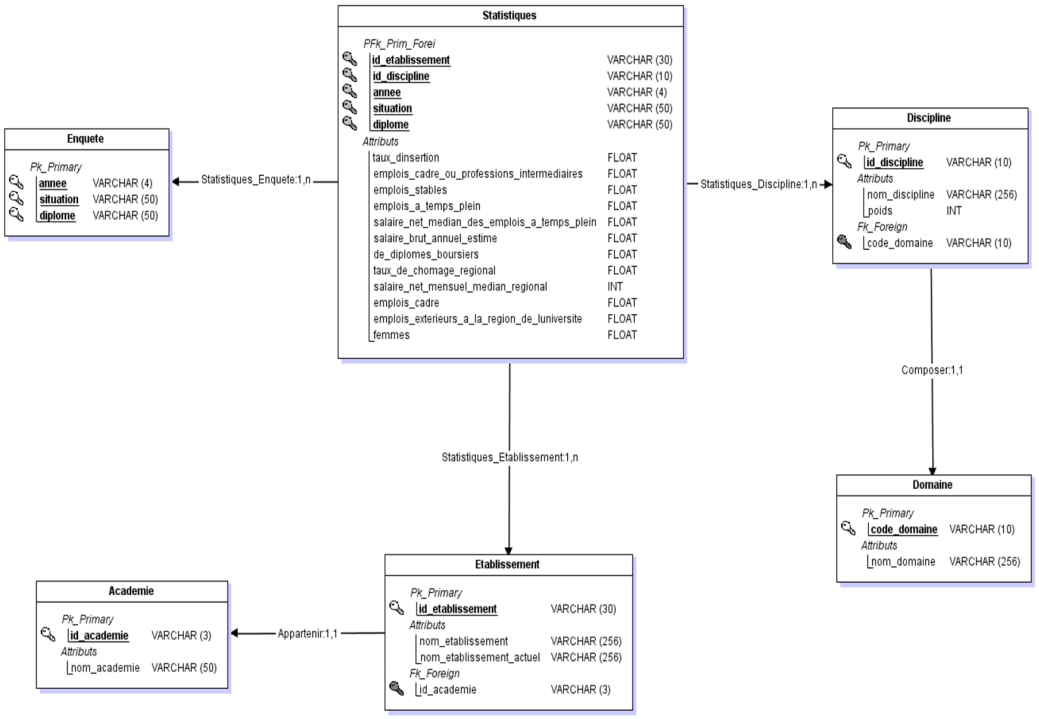

FIGURE 2 – Modélisation logique des données

la première chose qu'on peut remarquer c'est le fait que la relation **Statistique** a était transformer en une table avec des clés étrangères qui référencent les tables qui font partie de la relation. Ce qui veut dire que pour récupérer des statistiques précises il faudra préciser les clés primaires de l'enquête, de l'établissement et la discipline sur la quel les statistiques sont faites.

Nous pouvons également remarquer que des clés étrangères ont été ajoutées aux entités Discipline et Établissement, ce qui est totalement cohérent avec le MCD. En effet, l'ajout de ces clés garantit qu'un établissement appartient toujours à une académie spécifique et que nous pouvons donc avoir des statistiques selon les académies. Pour la discipline également, la clé étrangère ajoutée assure qu'une discipline appartient bien à un domaine.

3.3 Stockage des données

La modélisation logique étant faites, on a créé les tables associées avec le script SQL fournit par l'outil jMerise.
Pour l'insertion des données, on a implémenté un script python qui fait le prétraitement avec pandas, qui se connecte au serveur MySQL à l'aide MySQL.connector et qui insère les données dans chaque table en veillant à respecter les contraintes d'intégrités entre les tables.

4 Visualisation des données

La visualisation des données est faite à l'aide des frameworks seaborn et Maptlotlib de Python sous des Jupyter NoteBook.

5 Conclusion

Ce projet a été l'occasion de mettre en pratique sur un projet concret les notions et connaissances acquises durant le module. Nous avons réussi à modéliser, insérer et visualiser les données.
Nous aurions aimé proposer des services et une implémentation qui serait disponible sous une application web, mais malheureusement nous n'avons pas pu aller plus loin par manque de temps.