COUILLEROT Carol
MAHIER Loïc
PHALAVANDISHVILI Demetre

Rapport de projet *

*rapport réalisé sous LaTeX

Sommaire

1. Introduction

L'objectif de ce projet est de réaliser un entrepôt de données (OLAP) ainsi que des requêtes intéressantes sur un ou plusieurs jeu de données libres (open data). Pour ce faire nous avons choisis deux jeux de données : un sur les hébergements collectifs en France et l'autre sur les communes. Nous avons également choisis de réaliser ce projet en PL/SQL ainsi que d'utiliser Talend pour nettoyer nos données et concevoir nos tables relationnelles.

2. Choix des données

Nous avons trouver nos données sur le site "opendatasoft", elles sont également présente sur le site "data.gouv". Le premier jeu de données est sur les hébergements collectifs en France : c'est à dire les hôtels, les campings et les résidences avec des informations sur leur location, leur classement (nombre d'étoile) et leur capacité d'accueil notamment. Le deuxième jeu de donnée recueil toutes les communes de France, en indiquant leur population, leur superficie, leur code postal ainsi que leur département et leur région entre autre. Ce dernier nous permet d'affiner nos requête, d'en proposer des plus complexes mais aussi de pouvoir faire des regroupements et des classements par région et par département. Nous allons ainsi pouvoir faire des requêtes sur le classement (nombre d'étoile) de ces hébergements par département et région. Nous pourrons aussi regarder par commune, le nombre d'hôtels par habitant ou bien même faire une comparaison du nombre d'hébergement par région en fonction de l'année.

3. Constellation de fait

Après avoir choisis nos jeux de données, nous avons décidé de distinguer deux tables de faits : une propre aux hébergements avec des informations sur leur capacité d'accueil et leur classement par exemple. Et une seconde propre aux communes, avec leur population, leur superficie et leur location (département, région). Quatre dimensions s'y ajoutent : une pour les date, une pour

les adresses des hébergement, une pour les informations complémentaires des hébergements et enfin une relative aux communes avec un certain nombre de leur caractéristiques. Pour pouvoir faire des requêtes intéressantes et donc pour pouvoir joindre nos deux tables de faits, nous utilisons l'identifiant de l'hébergement qui est présent dans les deux tables de faits. Cela nous permet ainsi d'accéder aux caractéristiques propres à l'hébergement ainsi que celles propres à la commune de cette hébergement.

Pour des raisons pratiques nous avons également choisis de faire une vue précise sur la location. Celle-ci nous affiche pour chaque commune, son code postale, son code INSEE, son département et sa région. Le code INSEE est essentiel puisqu'il est unique, en effet ils se trouve que plusieurs communes ont le même code postale. Cette vue simplifiera nos requêtes complexes, d'autant que nous l'utiliserons fréquemment.

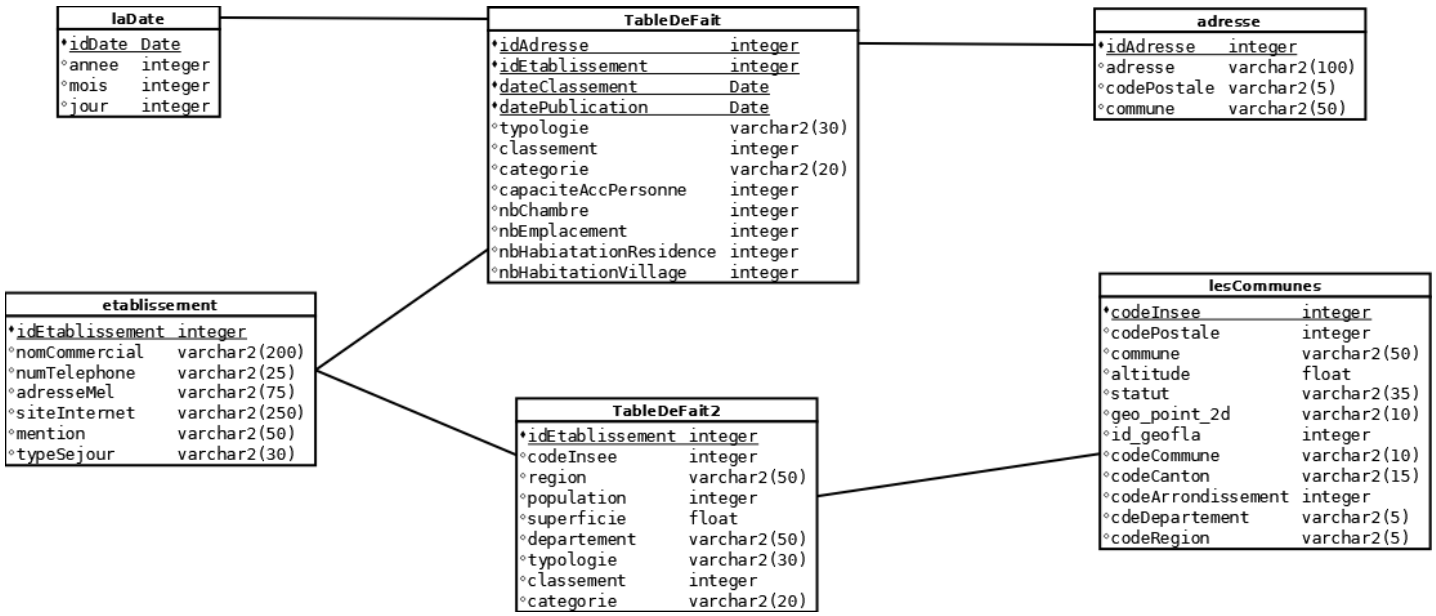

FIGURE 1 – La constellation de fait qui structure notre entrepôt de données

4. Intégration avec Talend

Après avoir défini notre schéma en étoile, notre prochaine objectif c'était de commencer à integrer nos dataset. Pour cela nous avons utiliser le logiciel Talend permettant de transformer nos données brut vers les tables relationnels de la base des données Oracle. Pour réaliser cette transformation nous avons eu plusieurs étapes intermédiaire. La première étape consisté à lire le fichier csv et le décomposer en plusieurs tables conforme au schéma en étoile définit dans la partie précédente. Pour cette décomposition nous avons utilisé à plusieurs reprises la fonction interne du logiciel Talend tMap.

La première point importante de l'utilisation de Talend, c'était la création du table laDate. Dans notre première dataset nous avons deux attributs date (datePublication et dateClassement),en utilisant les fonctions tUnite et tUniqRow de Talend, nous obtenons les dates uniques qu'on met dans le table laDate en ajoutant les attributs calculés annee,mois et jour.
La deuxième point importante de l'utilisation de Talend, c'est la création de la table lesCommunes, pour laquelle nous avons besoin d'utiliser les deux datasets pour fait l'intersection des codes postaux, pour cela au début nous avons utiliser tUniqrow pour avoir les communes uniques à partir de la première datasets puis nous avons fait l'intersection avec la deuxième dataset en utilisant tMap.
La troisième point importante de l'utilisation de Talend, c'est l'intégration des données nettoyer avec tMap dans la base des données Oracle. Pour cela nous avons utiliser la fonction tOracleOutput permettant inséré ou modifié les données à partir les tables défini dans tMap. Une fois configurer l'accès au base des données, on précise les noms des table(initialement vide) dans la base des donnée et a la fin on obtient la base des données rempli des informations.

5. Requêtes d'analyses

Une fois nos données sur Oracle, nous avons réalisé une dizaine de requête d'analyse en PL/SQL. Nous avons d'abord fait quelques requêtes sur la première table de fait, puis sur la deuxième avant d'en concevoir des plus complexes englobant les deux tables. Pour ces requêtes nous avons essayé de

réutiliser un grand nombre d'extension de SQL tels que ROLLUP, CUBE et GROUPING SET notamment.

...

6. Conclusion

...