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 jeux de données libres (open data). Pour ce faire nous avons choisi deux jeux de données, l'un sur les hébergements collectifs en France et l'autre sur les communes. Nous avons également choisi 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 trouvé nos données sur le site https://public.opendatasoft.com/explore, elles sont également présentes sur le site https://www.data.gouv.fr. 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'étoiles) et leur capacité d'accueil notamment. Le deuxième jeu de données recueille 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 requêtes, 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'étoiles) 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'hébergements par région en fonction de l'année.

3. Constellation de fait

Après avoir choisi 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 (dépar-

tement, région). Quatre dimensions s'y ajoutent : une pour les dates, une pour les adresses des hébergements, une pour les informations complémentaires des hébergements et enfin une relative aux communes avec un certain nombre de leur caractéristique. Pour pouvoir faire des requêtes intéressantes et donc pour pouvoir joindre nos jeux de données, 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 d'un hébergement ainsi que celles propres à la commune de cet hébergement.

Pour des raisons pratiques nous avons également choisi 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 il se trouve que plusieurs communes ont le même code postal. Cette vue simplifiera nos requêtes complexes, d'autant que nous l'utiliserons fréquemment.

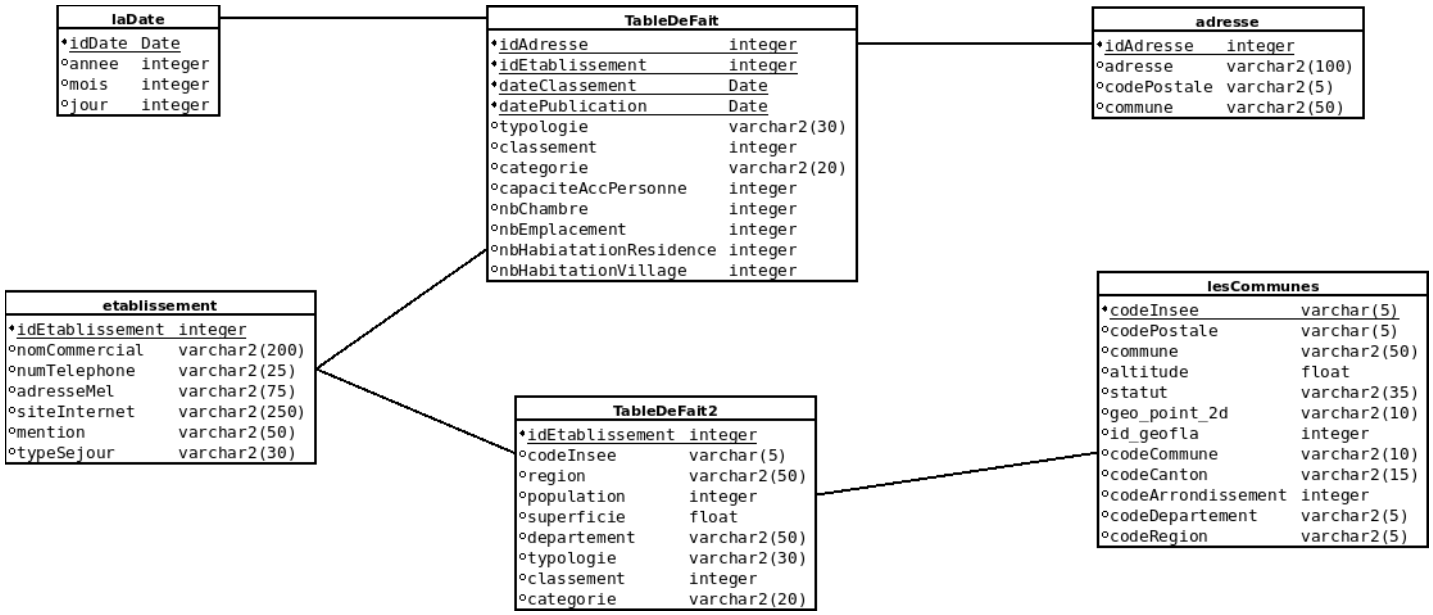

FIGURE 1 – La constellation de fait qui structure notre entrepôt de données

4. Intégration avec Talend

Après avoir défini notre constellation de fait, notre prochain objectif était de commencer à intégrer nos jeux de données. Pour cela nous avons utilisé le logiciel Talend permettant de transformer nos données brutes vers les tables relationnelles de la base de données Oracle. Pour réaliser cette transformation nous avons eu plusieurs étapes intermédiaires. La première étape consistait à lire le fichier au format ".csv" et à le décomposer en plusieurs tables conformément au schéma de fait défini dans la partie précédente. Pour cette décomposition nous avons utilisé à plusieurs reprises la fonction interne tMap du logiciel Talend.

Le premier point important de l'utilisation de Talend était la création de la table laDate. En effet, dans notre premier jeu de données nous avons deux attributs de type date (datePublication et dateClassement). En utilisant les fonctions tUnite et tUniqRow de Talend, nous obtenons des dates uniques qu'on met dans la table laDate en ajoutant les attributs calculés année, mois et jour.

Le deuxième point important de l'utilisation de Talend est la création de la table lesCommunes, pour laquelle nous avons besoin d'utiliser les deux jeux de données pour faire l'intersection des codes postaux. Pour cela nous avons utilisé tUniqrow afin d'obtenir des communes uniques à partir du premier jeu de données, puis nous avons fait l'intersection avec le deuxième jeu de données en utilisant tMap.

Le troisième point important de l'utilisation de Talend est l'intégration des données nettoyées avec tMap dans la base de données Oracle. Pour cela nous avons utilisé la fonction tOracleOutput permettant d'insérer et de modifier les données à partir des tables définies dans tMap. Une fois configuré l'accès à la base de données, on précise les noms des tables (initialement vides) dans la base de données et enfin on obtient l'entrepôt Oracle.

5. Requêtes d'analyses

Une fois nos données sur Oracle, nous avons réalisé une dizaine de requêtes 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.

Pour ce faire, nous avons identifié différents acteurs qui pourraient être amenés à utiliser ce jeu de données, par exemple les collectivités, un client lambda et un promoteur. Et c'est en nous mettant à leur place que nous avons essayé de faire le tour d'horizon des requêtes utiles : ainsi un client pourrait souhaiter savoir quel est le meilleur hôtel dans le département ou la région de ses vacances. De même un promoteur cherchant à bâtir un nouvel établissement (hôtel, résidence, camping, ...) aimerait savoir quel département est en manque d'hébergement. Ainsi on peut imaginer différents types de requêtes, les unes tournées vers l'analyse "touristique" informant sur le classement, le meilleur service, la proximité et les autres tournées sur l'analyse de "couverture", informant ainsi davantage sur la quantité d'hébergement dans une zone, la capacité d'accueil des hôtels dans une ville , etc.

Aussi, nous avons jugé utile de créer quelques vues, notamment une faisant la jointure entre les adresses de nos hébergements et la dimension lesCommunes, ce qui simplifie nos requêtes. Nous avons également fait des vues pour avoir de pré-calculé toutes les sommes ou moyennes qui pourraient nous servir dans nos requêtes. Enfin, pour améliorer les performances, nous avons choisis d'indexer les attributs annee, typologie,categorie,classement par un index bitmap : ils prennent en effet peu de valeurs, sont très fréquemment accédés et utilisés dans la clause where avec les opérateurs OR et AND. Enfin nous avons choisis d'ajouter deux index bitmap join, portant sur TableDeFait.dateClassement - laDate.annee ainsi que TableDeFait2.codeInsee - lesCommunes.codeInsee afin d'accélérer ces jointures.

6. Conclusion

Ainsi, nous avons réussi à créer notre premier entrepôt de donnée. Nous avons également réussi à proposer des requêtes pertinentes, analysant bien nos données. Ce projet nous aura permis de nous familiariser davantage avec les extensions de SQL tels que ROLLUP et CUBE notamment. La partie la plus ardue aura finalement été l'intégration des données avec Talend qui est un logiciel assez difficile à prendre en main mais au final extrêmement pratique et puissant. L'intégration ayant été faite sur Talend, tout en sachant que l'on peut directement envoyer nos données sur Oracle depuis Talend, il serait donc assez facile aujourd'hui d'ajouter de nouveaux jeux de données à notre entrepôt pour étendre notre champs d'analyse ou l'affiner encore davantage.