

COUILLEROT Carol
MAHIER Loïc
PHALAVANDISHVILI Demetre

Rapport de projet *

*rapport réalisé sous L^AT_EX

Sommaire

1	Introduction	3
2	Choix des données	3
3	Constellation de fait	3
4	Intégration avec Talend	5
5	Requêtes d'analyses	6
6	Conclusion	6

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 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 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 hébergements, 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.

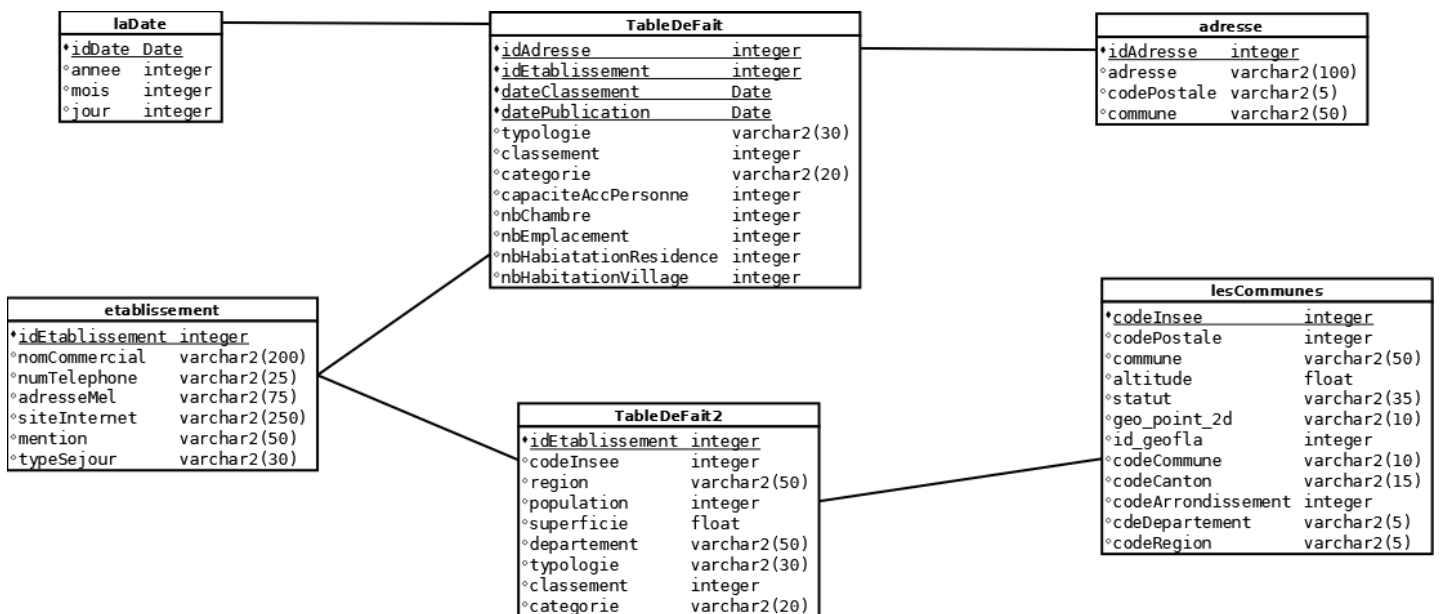


FIGURE 1 – La constellation de fait qui structure notre entrepôt de données

4. Intégration avec Talend

Après avoir défini notre constellation de fait, notre prochaine 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 conformes au schéma de fait défini dans la partie précédente. Pour cette décomposition nous avons utilisé à plusieurs reprises la fonction interne du logiciel Talend `tMap`.

Le premier point important de l'utilisation de Talend était la création de la table `laDate`. Dans notre premier jeu de données nous avons deux attributs 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 ou 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 la base de données remplie 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

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.