TP Chargement de données et révision du langage SQL en tant que LDD/LMD

1. Objectifs

L'objet du TP 1 est de se doter d'une volumétrie de données suffisante pour pouvoir ensuite conduire des tests de performance significatifs lors de prochains TPs. Les sous-objectifs du TP sont de deux natures.

1. Il s'agit d'une part d'explorer les mécanismes (module de chargement SQL Loader) qui permettent de charger les données provenant d'un fichier au format tabulé (CSV pour Comma Separated Value) au sein de tables d'une base de données relationnelle.
2. Il s'agit d'autre part de réviser vos acquis en matière de définition et de manipulation d'ordres SQL

2. Construction du schéma de base de données

2.1 Schéma de la base de données Communes

Le schéma relationnel de la base vous est donné, vous créerez les tables associées. Les attributs portant les contraintes de clés primaires sont en gras (la relation Commune est décrite sans clé primaire pour l'instant). Les contraintes de clés étrangères vous sont données sous la forme de contraintes d'inclusion. Les types des attributs vous sont également indiqués.

- Region(**reg varchar(4)**, chef_lieu varchar(46), nom_reg varchar(30))
- Departement(reg varchar(4), **dep varchar(4)**, chef_lieu varchar(46), nom_dep varchar(30))
  avec Departement(reg)⊆Region(reg)
- Commune(reg varchar(4),dep varchar(4),com varchar(4),article varchar(4),nom_com varchar(46),longitude float,latitude float,pop_1975 float,pop_1976 float, pop_1977 float,pop_1978 float,pop_1979 float,pop_1980 float,pop_1981 float,pop_1982 float,pop_1983 float,pop_1984 float,pop_1985 float,pop_1986 float,pop_1987 float,pop_1988 float,pop_1989 float,pop_1990 float,pop_1991 float, pop_1992 float,pop_1993 float,pop_1994 float,pop_1995 float,pop_1996 float,pop_1997 float,pop_1998 float,pop_1999 float,pop_2000 float,pop_2001 float,pop_2002 float,pop_2003 float,pop_2004 float,pop_2005 float,pop_2006 float,pop_2007 float,pop_2008 float,pop_2009 float,pop_2010 float)
  avec Commune(reg)⊆Region(reg)
  avec Commune(dep)⊆Departement(dep)

Un modèle conceptuel vous est donné qui retranscrit les entités considérées dans le schéma.

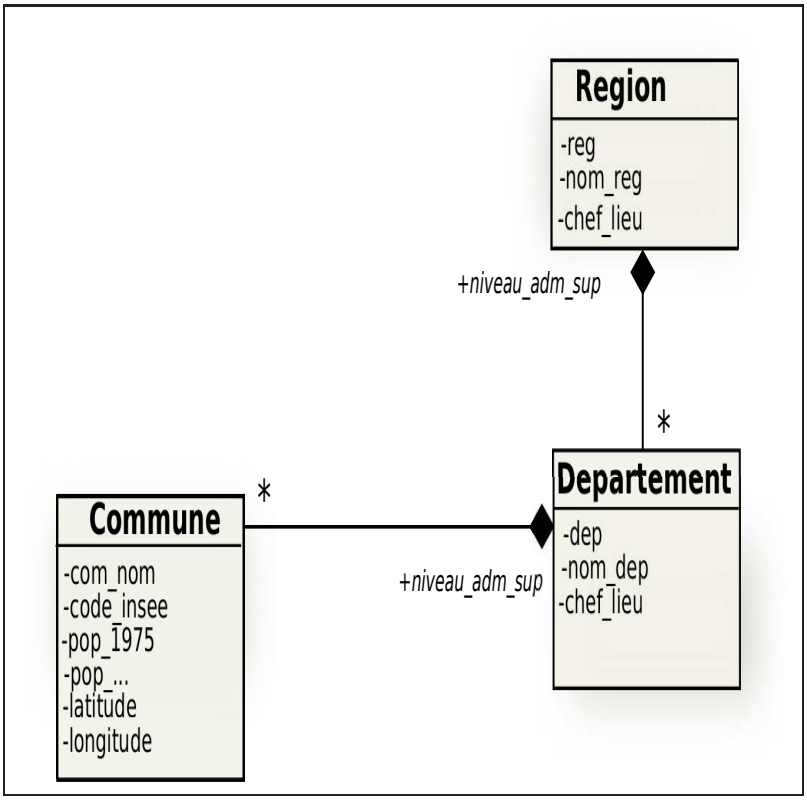

FIGURE 1 – Diagramme de classes UML

2.2 Construction de l'ensemble des tables et chargement

L'alimentation se fera au travers de l'utilitaire de chargement Oracle nommé SQL Loader. Vous disposez, pour ce faire, de différents fichiers de données au format tabulé [1]. Les données sur les communes ne concernent pour l'instant que la France métropolitaine (et la Corse). Par contre, les données sur les régions et départements portent également sur les ROM-COM. Deux fichiers de contrôle vous sont donnés sur les trois à construire (le fichier de contrôle Departement_ctl.txt reste à construire). Ces fichiers de contrôle vous permettent de définir les modalités d'insertion des données dans les tables. Un fichier archive vous est fourni avec un script SQL précisant l'ordre et les ordres SQL nécessaires à la définition et au chargement de l'ensemble du schéma. Les fichiers de données et les fichiers de contrôle de chargement sont également présents dans l'archive. Un schéma informel de chargement des données au travers de SQL Loader vous est fourni.

1. Crédits données : `www.insee.fr` et `http://freakonometrics.hypotheses.org/`

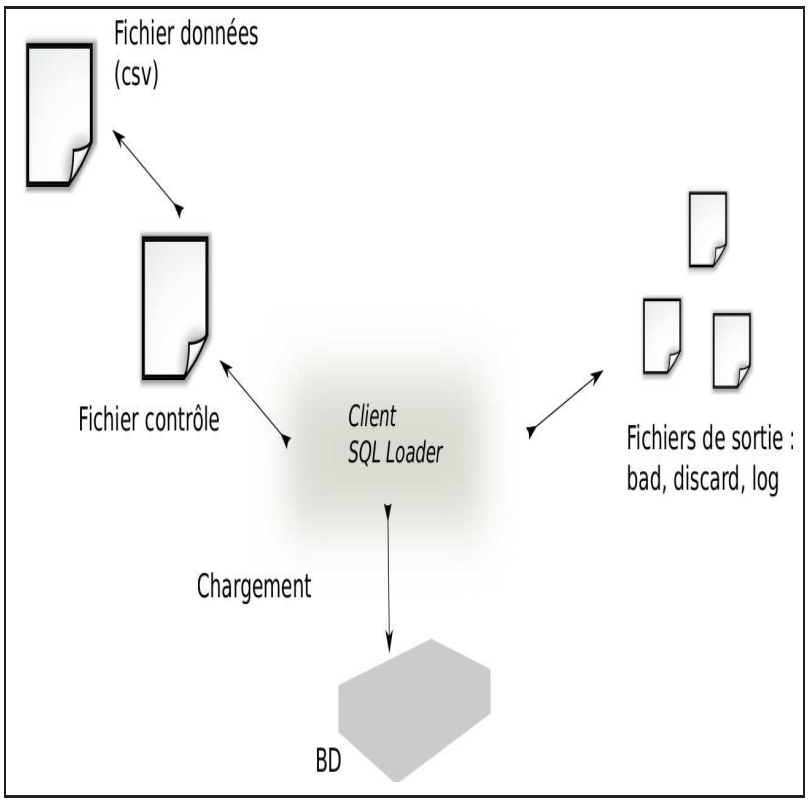

FIGURE 2 – Diagramme illustrant le module SQL Loader

A titre d'illustration, un fichier de contrôle est présenté ci-dessous. Il permet de charger des données provenant du fichier Region.csv dans la table Region. Le caractère de séparation dans le fichier tabulé est la tabulation ici représentée par X'9'

```
LOAD DATA
CHARACTERSET UTF8
INFILE 'Region.csv'
APPEND
INTO TABLE REGION
FIELDS TERMINATED BY X'9'
(reg  "to_char(:reg)",
chef_lieu "to_char(:chef_lieu)",
nom_reg "to_char(:nom_reg)"
)
```

L'appel à sql loader en ligne de commande est de la forme (serveur de données `venus` et base de données `master`) :
`sqlldr userid=nomUtilisateur/mdpUtilisateur@venus/master control=fichier_ctl.txt`

3. Evolution du schéma

Différentes mises à jour sont à appliquer sur le schéma (syntaxe ALTER TABLE) de manière à

le rendre le plus efficace et concis possible. Certaines mises à jour de données (syntaxe UPDATE nom_table) sont également demandées.

Table Commune

- Vous rajouterez une colonne nommée code_insee de type chaîne de caractères de longueur 6 à la table Commune
- Vous mettrez à jour le contenu de cette colonne qui prendra pour chaque tuple, la valeur obtenue par concaténation de dep et com
- Vous définirez une contrainte de clé primaire portant sur cet attribut code_insee
- Le schéma de la base de données comprend une redondance (attribut reg) qu'il convient de supprimer par souci d'efficacité. Cette redondance provient du contenu du fichier tabulé des données sur les communes exploité.

4. Manipulation du schéma

Vous définirez les requêtes suivantes en SQL. Vous pouvez connaître le temps nécessaire à l'exécution (+ restitution des données) d'une requête à l'aide de la variable d'environnement de sqlplus : `timing` (set timing on ou off). L'ensemble des variables d'environnement disponibles au sein du module sqlplus est listé au travers de la commande `show all`.

1. Nom et population en 2010 des communes de l'Hérault,
2. Donner les communes qui sont à la fois chef lieu de région et de département
3. Donner les communes qui sont chef lieu de département sans être chef lieu de région
4. Donner les communes qui sont chef lieu de région sans être chef lieu de département
5. Donner les communes qui dépendent du même chef lieu de département que FLORENSAC (code INSEE 34101)
6. Donner le nombre de départements par région
7. Donner le nombre de communes et de départements par région
8. Donner pour la plus petit commune (ou les plus petites communes) de France métropolitaine (en nombre d'habitants en 2010), son nom, le nom de son département et le nom de sa région
9. Ecrivez une requête qui permet de faire ressortir le manque d'informations pour les communes appartenant aux régions de l'ultra-marin (GUYANE ou MAYOTTE par exemple)
10. Nom, populations en 1975 et en 2010 des communes qui ont connu un afflux de population entre 1975 et 2010
11. Nom, population en 1975 et population en 2010 des communes de Languedoc-Roussillon (triées par ordre alphabetique) qui ont connu une diminution de population entre 1975 et 2010
12. Vue qui collecte les informations des communes (nom commune, nom département, nom région et recul de population) dont la population a baissé entre 2000 et 2010 puis consultation de cette vue
13. A partir de la vue, existe t'il des départements dont aucune commune n'a subi de recul de population entre 2000 et 2010 ?
14. Nom de la ou des communes du Languedoc-Roussillon qui ont le plus grand différentiel en terme de diminution de population (nombre de personnes) entre 1975 et 2010