Cancel Paul
Leroy Quentin
IUT-Info Groupe C

Rapport SAE 2.04
Base de Donnée

Exploitation d’une base de données
avec l’exemple des Jeux Olympiques.

Informations initiales :

En 1894, le baron français Pierre de Coubertin décide de remettre au goût du jour les jeux olympiques de la Grèce antique. La première édition se déroule en 1894 à Athènes. Au début réservés à des disciplines classiques, les jeux s'étendent à partir de 1924 aux disciplines hivernales. Lors de la création des Jeux olympiques d'hiver en 1924 et jusqu'en 1992, les Jeux d'été et d'hiver sont organisés tous les 4 ans, la même année. Depuis, chacun se déroule avec un décalage de 2 ans.
Sur le site Kaggle des contributeurs ont collecté toutes les participations de tous les athlètes à toutes les épreuves de tous les jeux jusqu'aux jeux de Rio en 2016.

Exercice 1 : Utilisation des commandes Unix pour gérer les données.

1. Combien y a-t-il de lignes dans chaque fichier ?
   Pour le fichier athlete_events.csv :
   wc -l athlete_events.csv
   271117 lignes

   Pour le fichier noc_regions.csv :
   wc -l noc_regions.csv
   0 lignes mais quand on lance le fichier dans nano, il y a 231 lignes

2. Afficher uniquement la première ligne du fichier athlète ?
   head -n 1 athele_events.csv

   "ID","Name","Sex","Age","Height","Weight","Team","NOC","Games","Year","Season","City","Sport","Event","Medal"

3. Quel est le séparateur de champs ?
   Le séparateur de champs est : ","

4. Que représente une ligne ?
   Une ligne du fichier correspond aux informations sur un athlète (âge, poids, taille, équipe et noc) et à quelle épreuve il a participé, la date de l'épreuve, et s'il a gagné une médaille

5. Combien y a-t-il de colonnes ?
   head -n 1 athlete_events.csv | tr “,” “\n” | wc -l
   15 colonnes dans le fichier athlete_events.csv.

6. Quelle colonne distingue les jeux d’été et d'hiver ?
   C’est la colonne « Season » qui distingue les jeux d’été et les jeux d’hiver
   Les jeux d’été sont notés « Summer » et pour ceux d’hiver, « Winter »

7. Combien de lignes font référence à Jean-Claude Killy ?
   grep “Jean-Claude Killy” athlete_events.csv | wc -l
   6 lignes font reference à Jean-Claude Killy

8. Quel encodage est utilisé pour ce fichier ?
   file -i athlete_events.csv
   Le fichier athlete_events.csv est encodé en us-ascii

9. Comment envisagez-vous l’import de ces données ?
   Avec la création d’une table temporaire “import” possédant l’intégralité des colonnes que possède le fichier puis on utilise cette commande pour importer:

   \copy import from athlete_events.csv delimiter ',' HEADER csv NULL AS 'NA';
   “NULL AS NA” permet de passer tous les « NA » à NULL pour nous faciliter l’import de ces données.

Exercice 2 : Import des données dans une fichier SQL.

1. On commencera par créer une table import avec en clés : id, name, sex, age, height, weight, team, noc, games, year, season, city, sport, event, medal.

2. Puis on utilise la commande précédemment vue qui copiera l'intégralité des données du fichier csv à l'intérieur de la table temporaire import.

3. On utilise la requête suivante pour supprimer toutes les données correspondantes :
   DELETE FROM import WHERE sport = 'Art Competitions' OR year<1920;

4. On crée maintenant une table temporaire noc avec comme colonnes noc et pays. Puis on utilise la requête
   \copy import from noc_regions.csv delimiter ',' HEADER csv

Pour que le script soit idempotent on fait bien attention à inclure un DROP TABLE de chacune des tables que l'on crée dans le fichier d'importation.

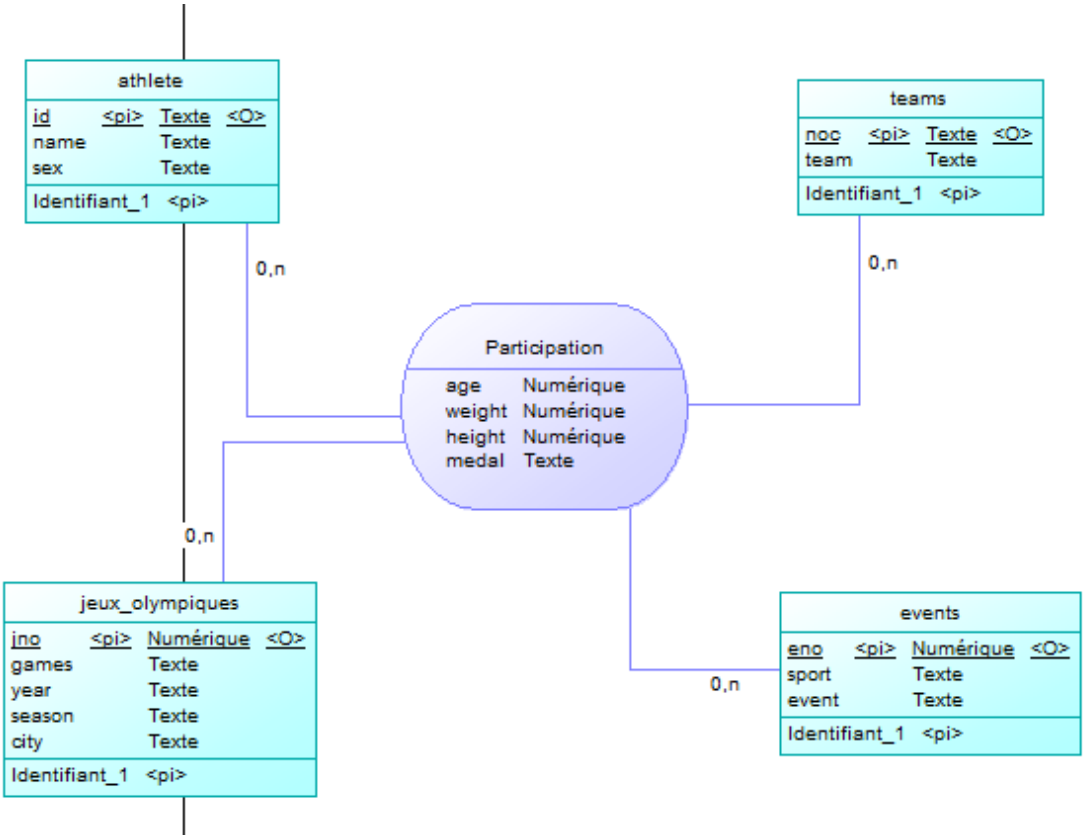
Exercice 3 : Requêtage sur les fichiers de départ.

Les réponses à cet exercice sont placées dans le fichier “requetes.sql”. Avec un commentaire pour le numéro de la question et un commentaire pour le renvoi de la requête.

Exercice 4 : Ventilation des données

La ventilation des données est un moyen que nous avons d’éviter des redondances dans le fichier et qui nous permet de diviser la taille d’un fichier en plusieurs tables comportant au total les mêmes informations pour au final moins d’espaces consommés.

Un Modèle Conceptuel de Donnée qui serait plus adapter à notre base de données pourrait ressembler à quelque chose comme ceci où l’on divise le fichier import en une table d’athlète, une table d’équipes, une table de jeux olympiques organisés, une table pour les différents évènements et toutes ces tables sont reliées par l’intermédiaire de la liaison Participations.



Avec un Modèle logique de donnée correspondant à ceci :

athlete : id , name, sex
teams : noc , team
jeux_olympiques : jno , games, year, season, city
events : eno , sport, event
participations : #id, #eno, #jno, #noc, medal, age, height, weight

On ajoute maintenant au fichier importation.sql les nouvelles tables ainsi que les requêtes permettant de les remplir tout en prenant soin d'ajouter chaque nouvelle table dans le DROP TABLE pour que le script SQL reste idempotent.

Tailles :
On utilisera la commande "SELECT pg_size_pretty(pg_relation_size('[table]'))" pour obtenir la taille de [table]

1) Taille : 41,5Mo

2) Taille : 46 Mo

3) Taille : 22.5 Mo

4) Taille : 10.8 Mo (Pour celui-ci il ne faut pas oublier de copier toutes les valeurs des différentes tables dans des fichiers csv pour avoir accès à leurs tailles une fois exportés).

Exercice 5 & 6 : Requêtage et personnalisation du Rapport.

Les requêtes créées pour l'exercice 5 sont toutes placées dans le fichier requetes.sql avec un commentaire pour les numéros des questions.

Pour la personnalisation, nous avons choisi la France en tant que Pays et la Gymnastique en tant que sport et nous avons imaginé ces différentes requêtes :

1. Médaillé d'argent ou de bronze français en gymnastique par année avec le nombre de médailles.
2. Gymnastes français de plus de 30 ans n'ayant jamais remporté de médaille
3. Nombre de gymnaste français ayant été médaillé
4. Moyenne des poids des gymnastes français n'ayant jamais été médaillé par sexe arrondie à 2 virgule près