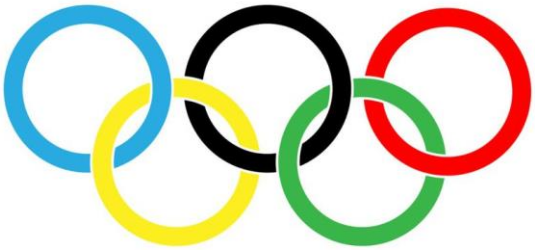
COUSIN Mathias, VANTARD Eddy – Groupe C

S2.04 – Exploitation de base de données

Données tirées des sessions de Jeux Olympiques de 1894 à 2016 comprises



BUT1 Informatique 2022/2023 – S2

IUT | Université de Lille

Au sein de ce rapport, il va être décrit et expliqué nos démarches lors de ce projet d'exploitation d'une base de données. Cela à avant tout consisté à importer, ventiler, analyser et requêter les différentes données que l'on nous a assigné.

Compréhension des données (Exercice 1) :

Les réponses aux questions qui suivent sont composées de la commande Unix permettant de trouver la réponse ainsi que de la réponse en question.

1) Nombre de lignes dans chaque fichier :

wc -l athlete_events.csv

271117 lignes

wc -l noc_regions.csv

231 lignes

2) Première ligne du fichier athlète :

head -n 1 athlete_events.csv

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

3) Séparateur de champs du fichier :

On observe grâce à la réponse précédente que c'est la virgule "," qui est utilisée comme séparateur de champs.

4) Représentation d'une ligne du fichier :

Encore grâce à la réponse à la question 2, on remarque qu'une unique ligne du fichier correspond à une participation d'**un** sportif à **une** épreuve lors d'**une** session de Jeux Olympiques accompagnées d'informations complémentaires.

Ainsi, le même athlète apparaît sur autant de lignes dans le fichier que de fois qu'il à participé à des épreuves différentes des Jeux Olympiques.

5) Nombre de colonnes :

On peut compter 15 colonnes pour ce fichier.

6) Colonne distinguant période estivale ou hivernale :

La onzième colonne prénommée "Season" dans la première ligne, désigne si la participation s'est déroulée lors de jeux d'hiver ou d'été.

7) Nombre de lignes faisant référence à Jean-Claude Killy :

cat athlete_events.csv | grep "Jean-Claude Killy" | wc -l

6 lignes

8) Encodage utilisé pour ce fichier :

file -i athlete_events.csv

Le fichier est encodé en "CSV text" et le charset en us-ascii.

9) Comment envisageons-nous l'importation de ces données ?

Il va être question de bien choisir les types de données à associer aux valeurs lors de la création d'une table temporaire import.

On utilisera alors la commande :

\copy import from athlete_events.csv WITH (delimiter ',', null 'NA', format CSV) ;

Enfin, nous ventilerons les données grâce à un MCD cohérent.

Importation des données (Exercice 2) :

Avant l'importation, il est nécessaire de créer une table temporaire permettant l'accueil des données. Ainsi, nous allons ici définir les types que nous avons choisi d'utiliser pour chaque donnée.

On remarque que la première colonne, correspondant à l'ID, se limite au final à 6 caractères au maximum dans le fichier donc nous décidons d'utiliser un CHAR(6).

Il en est d'ailleurs de même pour la quinzième colonne attribuant une médaille ou non à la performance étant donnée que cette valeur ne peut être que nulle, "Gold", "Silver" ou "Bronze". On utilise donc également un type CHAR(6).

La valeur déterminant le sexe de l'athlète ne peut valoir que 'M' ou 'F'. On utilise un CHAR(1).

L'âge, la taille ainsi que l'année sont trois valeurs que l'on associe à des type INT.

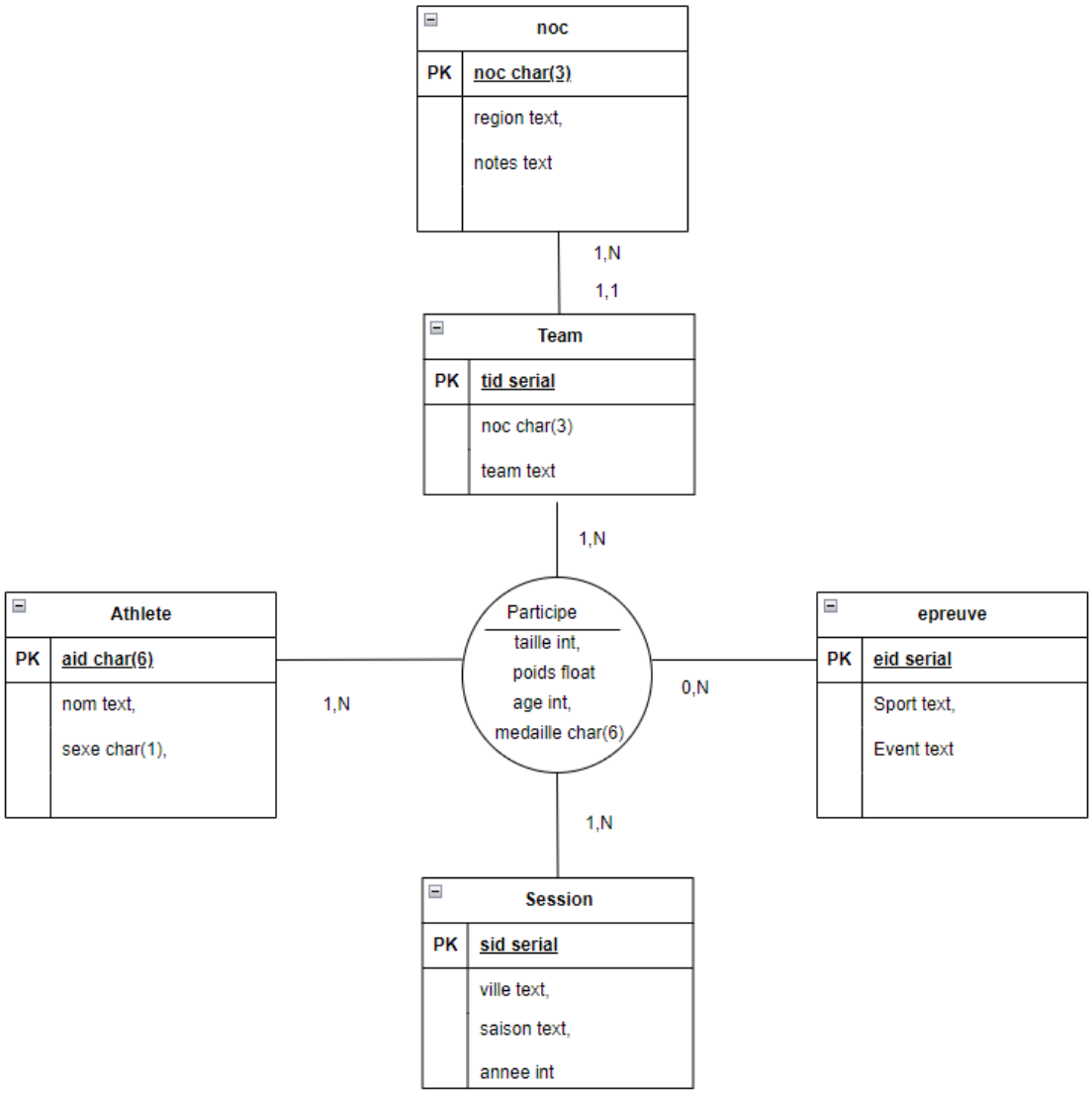
Le poids en revanche est en partie représenté avec des décimales dans le fichier donc on utilise le type FLOAT pour ce dernier.

Le NOC du pays représenté est un sigle toujours composée de trois lettres, le type CHAR(3) est alors le plus cohérent.

Pour finir, les colonnes restantes que sont "Name", "Team", "Games", "Season", "City", "Sport", "Event" sont toutes des chaînes de caractères trop longues ou trop imprévisibles pour pouvoir les limiter convenablement. On alors usage d'un type TEXT.

Ventilation des données (Exercice 4) :

Ci-dessous, le MCD correspondant à notre structuration :



La conception de ce modèle est avant tout dirigé par la réflexion suivante : certaines valeurs, telles que la taille, le poids, l'âge et l'obtention d'une médaille ou non est susceptible à des changements d'une année à l'autre et il est alors nécessaire de permettre à ces données de varier.

Ainsi, une quaternaire est présente au milieu du MCD possédant ces attributs variables. Cette dernière est liée à quatre tables aux données fixes :

- Athlète possède un identifiant, le nom et le sexe d'un athlète.

- Session correspond à une période de Jeux Olympiques et est alors définie par un numéro automatique, la ville accueillant les jeux, la saison et enfin l'année.

- Epreuve représente un type d'épreuve que l'on définit par un numéro automatique, la discipline correspondante ainsi que le nom de l'épreuve en particulier.

- Team complète cette quaternaire en possédant un numéro automatique, le nom de l'équipe et le noc du pays qu'elle représente.

Enfin, la table Team est reliée par le noc du pays représenté à la table noc elle-même possédant le noc, le nom du pays et une possible note à son sujet.

MLD associé :

epreuve(eid serial, sport text, event text)

session(sid serial, ville text, saison text, annee int)

athlete(aid char(6), nom text, sexe char(1))

team(tid serial, #noc char(3), team text)

noc(noc char(3), region text, notes text)

participe(#eid int, #sid int, #aid char(6), #tid int, taille int, poids float, age int, medaille char(6))

Questions de tailles :

1) Taille en octet du fichier récupéré :

wc -c athlete_events.csv

41500688 octets

2) Taille en octet de la table import :

On entre cette requête SQL :

```
SELECT table_name,
pg_relation_size(table_schema || '.' || table_name) As Taille_donnees,
pg_total_relation_size(table_schema || '.' || table_name) As Taille_totale
FROM information_schema.tables WHERE table_schema = 'pg_temp_9'
ORDER BY Taille_totale DESC;
```

```
 table_name | taille_donnees | taille_totale
-----------------+--------------------+---------------
 import          |       48078848 |   48119808
```

La taille totale de la table import est donc de 48119808 octets.

3) Taille en octet de la somme des tables créées :

On entre cette requête SQL :

```
SELECT SUM(pg_relation_size(table_schema || '.' || table_name)) As Somme_Taille_donnees,
SUM(pg_total_relation_size(table_schema || '.' || table_name)) As Somme_Taille_totale
FROM information_schema.tables WHERE table_schema = 'eddyvantardetu'
ORDER BY Taille_totale DESC;
```

```
 taille_donnees | taille_totale
--------------------+---------------
       24846336  |   42426368
```

Ainsi, la taille totale que représente la somme des tables crées vaut 42426368 octets.

4) Taille en octet que fait la somme des tailles des fichiers exportés correspondant à ces tables :

Exportation des tables PSQL :

```
\copy team to 't_Team.csv' WITH (format CSV, delimiter ',', NULL 'NA');

\copy epreuve to 't_Epreuve.csv' WITH (format CSV, delimiter ',', NULL 'NA');

\copy session to 't_Session.csv' WITH (format CSV, delimiter ',', NULL 'NA');

\copy noc to 't_Noc.csv' WITH (format CSV, delimiter ',', NULL 'NA');

\copy participe to 't_Participe.csv' WITH (format CSV, delimiter ',', NULL 'NA');

\copy athlete to 't_Athlete.csv' WITH (format CSV, delimiter ',', NULL 'NA');
```

On additionne ensuite les tailles des données exportées :

(Dans un répertoire ne comportant que les fichiers en question) : wc -c *

```
 3737819 t_Athlete.csv
   28581 t_Epreuve.csv
    4003 t_Noc.csv
 7905146 t_Participe.csv
    1129 t_Session.csv
   14932 t_Team.csv
11691610 total
```

La somme vaut alors 11691610 octets.

Requêtage personnel (Exercice 6) :

Cette toute dernière partie du rapport sera consacrée au choix d'un pays et d'une discipline sur laquelle on appliquera quatre requêtes que l'on estime intéressantes ou utiles.

Nous avons choisi arbitrairement la Norvège comme pays et le canoë comme discipline.

La première requête vise à récupérer le nombre de médailles d'or, d'argent et de bronze que la Norvège a pu remporter en canoë durant l'intégralité des jeux.

La seconde requête calcule quant à elle l'âge, la taille et le poids moyen des athlètes canoéistes norvégiens.

La troisième requête permet d'extraire le nombre d'athlètes présents par année croissante ainsi que le nombre de médailles remportées par ces derniers (bronze, argent, or et au total) pour cette même année. Cette requête peut nous faire conjecturer l'impact d'avoir beaucoup d'athlètes sur les réussites.

Pour finir, la dernière requête calcule le pourcentage des médailles remportées par la Norvège en canoë par épreuve de la discipline afin de savoir en quelle type d'épreuve de canoë la Norvège performe le plus.

COUSIN Mathias, VANTARD Eddy – Groupe C