PETREMAND Shems, GUIOT Quentin - Groupe D

SAE S2.04: Exploitation d'une Base De Données

BUT 1 Informatique S2 2022/2023

Université de Lille

Vous trouverez dans ce rapport des explications concernant le travail effectué durant cette SAE ayant comme base de données les Jeux Olympiques de 1894 à 2016 inclus.Deux fichiers ont été fournis afin de réaliser ce projet “athlete_event.csv” et “nos_regions.csv”.

Première partie: Compréhension des données.(Exercice 1)

Les réponses de cette section ont été obtenues à l’aide de commandes Unix.

1.Combien y-a-t-il de lignes dans chaque fichier?

Le fichier athlete_event.csv contient 271117 lignes tandis que le fichier noc_regions.csv est composé de 231 lignes.

Commande effectuée: *wc -l <nomdufichier>*

2.Affichage de la première ligne du fichier athlète:

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

Commande effectuée: *head -n 1 athlete_events.csv*

3.Quel est le séparateur de champs?

Comme nous pouvons le voir avec l’affichage de la première ligne du fichier athlete_events.csv le séparateur utilisé entre chaque champ est une virgule.

4.Que représente une ligne?

Chaque ligne du fichier représente toutes les informations concernant un participant à une épreuve dans un sport et à une session des Jeux Olympiques.
Nous pouvons avoir comme preuve l’affichage de la question 2.

5.Combien y-a-t-il de colonnes?

Encore une fois grâce à la réponse de la question 2 en additionnant les différents champs nous comptons 15 colonnes pour le fichier athlete_events.csv.

6.Quelle colonne distingue les jeux d’été et d'hiver?

La colonne Season prend uniquement les valeurs Summer et Winter” ce qui permet de différencier les jeux d’été de ceux d'hiver.

7.Combien de lignes font référence à Jean-Claude Killy?

Dans le fichier athlete_events.csv, 6 lignes font références à Jean-Claude Killy, cette information a été obtenue à l’aide de la commande:
grep “Jean-Claude Killy” athlete_events.csv | wc -l

8.Quel est l'encodage utilisé pour ce fichier?

L’encodage de ce fichier est “CSV text”.
Commande utilisée: *file athlete_events.csv*

9.Comment envisagez vous l’import de ces données?

L’importation de ces données va nécessiter de reproduire une table identique qui nous servira d'intermédiaire pour les prochaines parties tout en s’assurant que cette dernière récupère l'entièreté de la base de données.

Deuxième partie: Importation des données.(Exercice 2)

Pour commencer avec l’importation nous avons créé une table “import” qui concentrera la totalité des données du fichier athlete_events.csv, pour chaque chaque colonne il est important de bien définir le type qui la caractérise.

Concernant le champs”ID” la valeur dans le fichier original est un nombre entiers on lui donne donc le type int, nous pouvons d’ailleurs faire de même pour les catégories “Age”,”Height” et “Year” qui renferment respectivement l'âge et la taille du sportif et l’année où s’est déroulée l’épreuve.

Pour ce qui est du nom, le nombre de caractères est différent pour tout le monde et après recherche le nom le plus long de ce fichier comporte 110 caractères, nous pouvons donc lui attribuer le type varchar(110).
D’autres varchar sont attribués pour les champs “Team”,”City”,”Sport”,”Event” et “Medal” où la valeur maximale donnée en paramètre correspond à la chaîne la plus longue pour chaque champ dans le fichier de base.

Pour terminer avec la table import, elle est composé de 3 char, le premier concerne le sexe des participants avec comme uniques valeurs “F” et “M” nous pouvons donc mettre le type à char(1), ensuite NOC permettant de reconnaitre les noms des pays est toujours composée de 3 charractères on lui attribue alors le type char(3) et enfin

avec season comme dit precedemment ne prend que les valeurs “Summer” et “Winter” toute deux composées de six lettres on lui applique ainsi le type char(6);

Pour ensuite effectuer l’importation il nous suffit de faire la commande ci-contre:

\copy import from athlete_events.csv with(format CSV,delimiter ',', NULL 'NA',HEADER);

Pour être en accord avec les consignes il ne nous reste plus qu’à enlever les lignes indésirables qui sont les années antérieurs à 1920 et tous les sports où le nom commencerait par “Arts”.

L’importation de la seconde table se fera de manière beaucoup plus rapide puisqu’il nous était demandé de l’import tel-quel sans aucune modifications, dans notre cas elle s'appellera import2.Nous pouvons une fois la table prête effectuer cette commande:

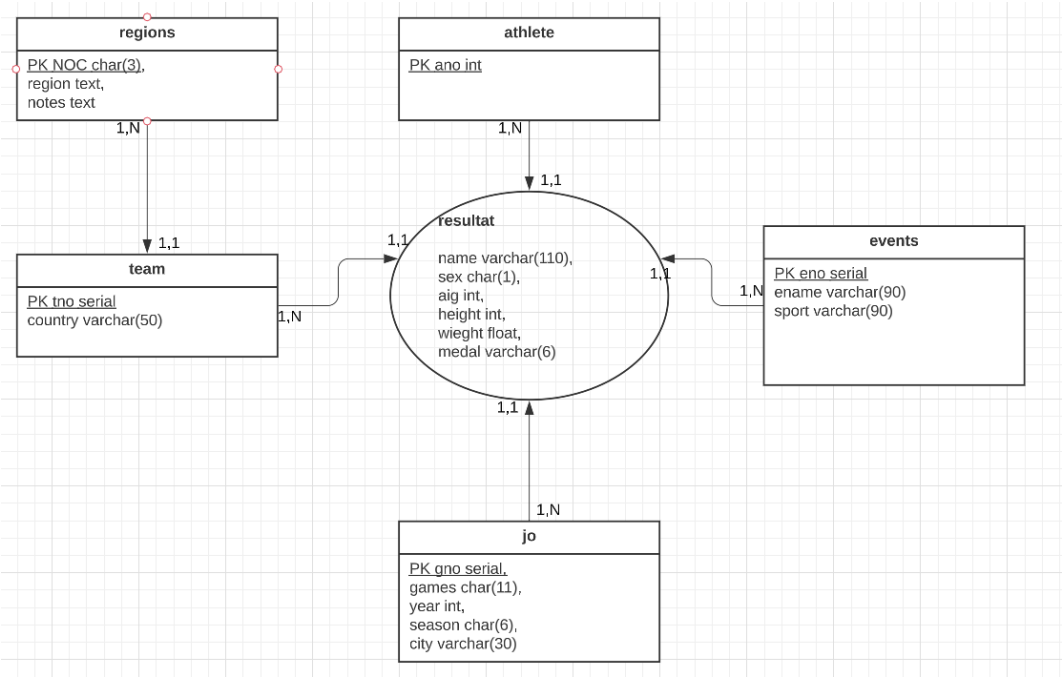
\copy import2 from noc_regions.csv with(format CSV,delimiter ',', NULL 'NA', HEADER);

Pour le bien de la suite du projet nous avons prit la décision d’effectuer une unique modification sur la table import2 pour réaliser de manière plus efficace correspondant à cette commande SQL:

UPDATE import2 SET noc = 'SGP' WHERE noc = 'SIN';

Troisième partie: Ventilation des données.(Exercice 4)

1.Vous trouverez ci-joint le MCD que nous avons réalisé.



Nous avons décidé de garder ici la table “regions” telle quelle afin de pouvoir transmettre à la table “team” le “NOC” qui s’accorde bien avec “country” afin de regrouper le pays et son noc.

Il y a ensuite la table “jo” qui indique que les Jeux Olympiques se déroulent dans une ville précise à une certaine saison et une certaine année.

Concernant la table “athlete” on n’y retrouve que le numéro de l'athlète afin de faire en sorte que l'athlète ne change pas d’ID même s’il change de catégorie ce qui explique notre choix de regrouper toutes les caractéristiques des athlètes dans la table résultat.

La table”events” quant à elle indique qu’une épreuve est caractérisée par son nom et par son sport.

Et enfin, la table résultat découle de la quaternaire avec comme clé primaire la réunion des quatre autres tables avec les caractéristiques de l'athlète et la potentiel médaille qu’il aurait gagné.

MLD correspondant:

athlete(ano int)

events(eno serial, ename varchar(90), sport varchar(30))

team(tno serial,country varchar(50), #NOC char(3))

regions(NOC char(3),region text,notes text)

jo(gno serial,games char(11),year int,season char(6),city varchar(30))

participe(#eno serial, #ano int, #gno serial, #tno serial, sex char(1),age int, height int, weight float, medal varchar(6))

2. Une question de taille!

-Quelle taille en octet fait le fichier récupéré?
Le fichier récupéré fait 41500688 octets.

Commande utilisé: *wc -c athlete_ events.csv*

Quatrième partie: Personnalisation du rapport.(Exercice 6)

Cette dernière partie de notre travail consistait à choisir arbitrairement un pays et une discipline et d'appliquer quatre requêtes à ces contraintes.

Nous avons donc choisi la Colombie comme pays et le weightlifting comme sport.

Notre première requête nous permet d'avoir le nom et le poids des colombiens ayant gagné un médaille d'or en faisant du weightlifting.

La seconde requête renvoie les noms des différentes colombiennes ayant participé à des épreuves de weightlifting.

La troisième requête permet de voir les noms des colombiens homme ayant remporté des médailles ou non dans les épreuves de weightlifting à partir de 2012.

Et pour finir finir avec la quatrième requête affichant le nom des épreuves de weightlifting recensées depuis 2010 où ont participé des colombiens.