SAé S2.04 : Exploitation de BDD

Objectif

L'objectif de cette SAé consiste à :

— Comprendre un ensemble de données réelles

— Savoir importer ces données

— Savoir normaliser les données ainsi récupérées

— Savoir réaliser des requêtes classiques, et des requêtes d'extraction de données pour exploitation statistique sur cet ensemble.

Travail à réaliser

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.

Ces données sont disponibles de manière brute (ficher plat) sur Moodle (et sur Kaggle bien sûr)

— `athletes_event.csv` contenant toutes les données sur chaque participation (y compris les médailles obtenues)

— `noc_regions.csv` contenant le décodage des codes pays.

Le travail de cette SAé consiste à importer, ventiler, analyser et requêter les différentes données récupérées sur ce site. Traiter ce problème ne se fait pas séquentiellement question après question. Cela nécessite quelques essais et tentatives avant d'avoir la bonne démarche. Il se peut que ce ne soit qu'après avoir avancé un peu que vous saurez répondre efficacement aux premières questions.

Exercice 1 : Comprendre les données

Q1. Analyse du fichier récupéré

Avant toute action sur la base de données, il est avant tout nécessaire d'analyser le fichier récupéré. Les commandes Unix vont vous aider ! Vous devez impérativement avoir répondu à ces questions sur cette feuille à la fin de la première séance, et avant toute action sur la base.

Les réponses à cet exercice seront décrites dans le rapport.

Sous Unix, avant une quelconque action sur la base de données, écrire `la commande Unix` qui permet de répondre aux questions suivantes:

1. Combien y-a t-il de lignes dans chaque fichier ? .....
2. Afficher uniquement la première ligne du fichier athlète .....

3. Quel est le séparateur de champs ? .....
4. Que représente une ligne ? .....
5. Combien y-a t-il de colonnes ? .....
6. Quelle colonne distingue les jeux d'été et d'hivers ? .....
7. Combien de lignes font référence à Jean-Claude Killy ? .....
8. Quel encodage est utilisé pour ce fichier ? .....
9. Comment envisagez vous l'import de ces données ? .....

Exercice 2 : Importer les données

Les réponses à cet exercice seront placées dans le script `importation.sql`. On considèrera que la commande d'exécution est lancée dans le même répertoire dans lequel sont placés les fichiers.

1. Créer une table `import` permettant l'importation de ces données
2. Remplir cette table avec les données récupérées
   S'assurer que les types de colonnes soient les plus restrictifs possibles (des `int` pour les colonnes contenant des entiers, des `char(x)` pour les données textuelles de taille `x` etc ...)
3. Supprimez toutes les données strictement inférieures à 1920 (qui semblent peu crédibles). On ne travaillera plus qu'avec des données à partir de 1920.
   Il doit vous rester 255.080 lignes dans la table import. On ne veut plus en perdre une seule !
4. Importer tel quel le fichier `noc_regions.csv`

Ce script doit être `idempotent`. On doit pouvoir le lancer autant de fois qu'on le souhaite.

Exercice 3 : Requêtage sur les fichiers de départ (import et noc)

Les réponses à cet exercice seront placées dans le fichier `requetes.sql`. On veillera à mettre un commentaire avec le numéro de la question.

Fournir les commandes ou les requêtes qui indiquent:

Q1. Combien de colonnes dans import ? (1 valeur)

Q2. Combien de lignes dans import ? (1 valeur)

Q3. Combien de codes NOC dans noc ? (1 valeur)

Q4. Combien d'athletes différents sont référencés dans ce fichier (1 valeur)

Q5. Combien y-a t-il de médailles d'or dans ce fichier ?(1 valeur)

Q6. Retrouvez Carl Lewis; Combien de lignes se réfèrent à Carl Lewis ? (1 valeur)

Exercice 4 : Ventiler les données

Q1. Normalisation des données

Décomposez la table `import` en plusieurs tables (3è forme normale : 3NF). Certaines tables sont évidentes afin d'éviter les redondances, d'autres permettent de séparer les données pour éviter les redondances.

1. Fournir le MCD correspondant à votre structuration.
2. Fournir le MDL associé.
3. Complétez le fichier `importation.sql` avec toutes les actions d'importation et de création/remplissage des différentes données.

   On fera en sorte que ce script soit *idempotent* (on peut le lancer autant de fois que l'on veut, il donne toujours le même résultat.)

Q2. Une question de taille !

Les réponses à cet exercice seront placées dans le rapport.

1. Quelle taille en octet fait le fichier récupéré ? .....
2. Quelle taille en octet fait la table import ? .....
3. Quelle taille en octet fait la somme des tables créées ? .....
4. Quelle taille en octet fait la somme des tailles des fichiers exportés correspondant à ces tables ? .....

Exercice 5 : Requêtage

Les réponses à cet exercice seront placées dans le fichier `requetes.sql`. On veillera à mettre un commentaire avec le numéro de la question.

Q1. Liste des pays classés par participation aux épreuves (2 cols)

Q2. Liste des pays classés par nombre de médailles d'or (2 cols)

Q3. Liste des pays classés par nombre médailles totales (2 cols)

Q4. Liste des sportifs ayant le plus de médailles d'or, avec le nombre (3 cols)

Q5. Nombre de médailles cumulées par pays pour les Jeux d'Albertville, par ordre décroissant (2 cols)

Q6. Combien de sportifs ont fait les jeux olympiques sous 2 drapeaux différents, le dernier étant la France ? (1 valeur) Selon vous quel est le plus connu/célèbre/titré/... ?

Q7. Combien de sportifs ont fait les jeux olympiques sous 2 drapeaux différents, le premier étant la France ? (1 valeur) Selon vous quel est le plus connu/célèbre/titré/... ?

Q8. Distribution des âges des médaillés d'or (2 cols)

Q9. Distribution des disciplines donnant des médailles aux plus de 50 ans par ordre décroissant (2 cols)

Q10. Nombre d'épreuves par type de jeux (hivers,été), par année croissante (3 cols)

Q11. Nombre de médailles féminines aux jeux d'été par année croissante (2 cols)

Exercice 6 : Personnalisation du rapport

Choisissez un sport et un pays.
Proposez 4 requêtes dessus (par exemple : les résultats, la participation, la place, les particularités, ...).

Le correcteur portera une attention particulière à la singularité de cet exercice.

Les requêtes seront placées dans le fichier `requetes.sql` et une section spéciale du rapport présentera votre analyse.

À rendre pour la partie BDD

Une archive zip déposée sur Moodle **le 10 avril minuit maxi** avec

1. Un rapport explicatif de 5 à 7 pages sous format PDF contenant:
   - une page de garde (titre, logos, noms des étudiants)
   - — une partie expliquant et commentant votre démarche pour chacune des étapes : compréhension (Ex1), importation (Ex2), ventilation (Ex4) des données ainsi que pour l'Ex6.
   - N'oubliez pas d'inclure votre MCD en tant qu'image dans le rapport.
2. Le fichier `importation.sql` qui permet de tout recréer à partir des fichiers fournis placés dans le même répertoire (avec Exo2 et Exo4). On doit pouvoir lancer ce script autant de fois qu'on le souhaite.
3. Le fichier `requetes.sql` qui permet d'exécuter vos différentes requêtes d'interrogation (Exo3, Exo5, Exo6). Vous mettrez un commentaire à chaque requête avec le numéro de l'exercice et la question correspondante.