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

Tous les ans, les étudiants voulant s'inscrire dans une formation du supérieur doivent passer par la plateforme nationale **ParcourSup**. Cette plateforme permet aux étudiants d'exprimer leurs choix de candidature à la poursuite d'étude et aux établissements d'indiquer les classés (en 2 phases maxi) puis les acceptés.
Le ministère de l'éducation nationale rend public les données agrégées (pas de données individuelles) issues de cet outil.

— Le site
— Les données

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

1. Combien y-a t-il de lignes ? Justifiez ! Il y a 13870 Lignes, on le vois grâce à la commande "wc -l parcoursup.csv
2. Que représente une ligne ? Une ligne correspond au données d'un établissement.
3. Combien y-a t-il de colonnes ? Justifiez ! Il y a au total 118 colonnes, on le vois grâce la commande : head -n 1 parcoursup.csv | tr ';' '\n' | wc -l
4. Quelle colonne identifie un établissement ? (numéro et nom de col) C'est la colonne D, soit la 4ème colonne avec le nom établissement
5. Quelle colonne identifie une formation ? (numéro et nom de col) C'est la colonne J qui correspond à "Filière de formation"
6. Combien de lignes font référence à notre BUT Informatique ? Il y a 1 ligne qui correspond au BUT, à la ligne 9849
7. Quelle colonne identifie un département ? (numéro et nom) C'est la colonne F qui correspond au "Département de l'établissement"
8. Comment envisagez vous importer ces données ? Nous créerons une table "import" pour mettre les données dans notre SQL
9. Quels problèmes identifiez vous dans ces données initiales ? (il y en a surement plusieurs, expliquez les clairement)
   Il y a beaucoup de redondance dans les données, certaines données comme des liens ne sont pas pertinentes.

Q2. Importer les données

Les noms des colonnes étant particulièrement complexes, on fera en sorte que l'ensemble des colonnes soit renommé à partir de n1.

1. Fournir un fichier `dico.xls` permettant la correspondance entre les numéros de colonnes et les noms du fichier initial. Expliquez comment vous vous y êtes pris pour le constituer.
2. Créer une table `import` permettant l'importation de ces données (fournir le code)
3. 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 ...)
4. Remplir cette table avec les données récupérées (fournir le code)
5. En s'appuyant sur la table `import` fournir les requêtes et les réponses qui permettent de savoir
   (a) Combien il y a de formations gérés par ParcourSup ? .....
   (b) Combien il y a d'établissements gérés par ParcourSup ? .....
   (c) Combien il y a de formations pour l'université de Lille ? .....
   (d) Combien il y a de formations pour notre IUT ? .....
   (e) Quel est le code du BUT Informatique de l'unversité de Lille ?
   (f) Citez 5 colonnes contenant des valeurs nulles

Exercice 2 : 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 par thèmes.

Il vous faut conserver suffisamment d'informations pour pouvoir faire les requêtes souhaitées et supprimer les redondances ou informations que vous jugez inutiles. On vous laisse le soin de filtrer les données pour ne conserver que les informations qui vous semblent les plus intéressantes (toutes les colonnes utilisées précédemment doivent au minimum être conservées).

1. Fournir le MCD correspondant à votre structuration
2. Ecrire le script `parcourssup.sql` qui permet de réaliser toutes les actions d'importation et de création/remplissage des différentes parcourssup.

   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 !

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 3 : Requêtage

À rendre dans un fichier `requetes.sql` avec commentaires afin de nous permettre d'identifier la question.
Nous partons du principe que les colonnes ont été numérotées à partir de `n1`

Q1. Ecrire une requête qui, à partir de `import` affiche le contenu de la colonne `n56` et le re-calcul de celle-ci à partir d'autres colonnes de `import` (2 cols).

Q2. Quelle requête vous permet de savoir que ce re-calcul est parfaitement exact ?

Q3. Ecrire une requête qui, à partir de `import` affiche le contenu de la colonne `n74` et le re-calcul de celle-ci à partir d'autres colonnes de `import` (2 cols).

Q4. Quelle requête vous permet de savoir que ce re-calcul est parfaitement exact ?

Q5. Ecrire une requête qui, à partir de `import` affiche le contenu de la colonne `n76` et le re-calcul de celle-ci à partir d'autres colonnes de `import` (2 cols). A partir de combien de décimales ces données sont exactes ?

Q6. Fournir la même requête sur vos tables ventilées

Q7. Ecrire une requête qui, à partir de `import` affiche la `n81` et la manière de la recalculer. A partir de combien de décimales ces données sont exactes ?

Q8. Fournir la même requête sur vos tables ventilées

À rendre pour la partie BDD

Une archive zip déposée sur Moodle **le 20 avril maxi** avec

1. Un rapport explicatif en PDF avec une page de garde (titre, logos,noms des étudiants) ainsi que le MCD, la réponse aux différentes questions de ce document (notamment (mais pas uniquement) Exo1 Q1, Exo1 Q2, Exo2 Q2 et Exo 3) et les commentaires que vous jugerez necessaires.
2. Le fichier `dico.xls` qui contient les correspondances de noms de colonnes
3. Le fichier `parcourssup.sql` qui permet de tout recréer
4. Le fichier `requetes.sql` qui permet d'exécuter vos différentes requêtes.