Université de Lille

R2.06 - BDD
BUT Info – N2

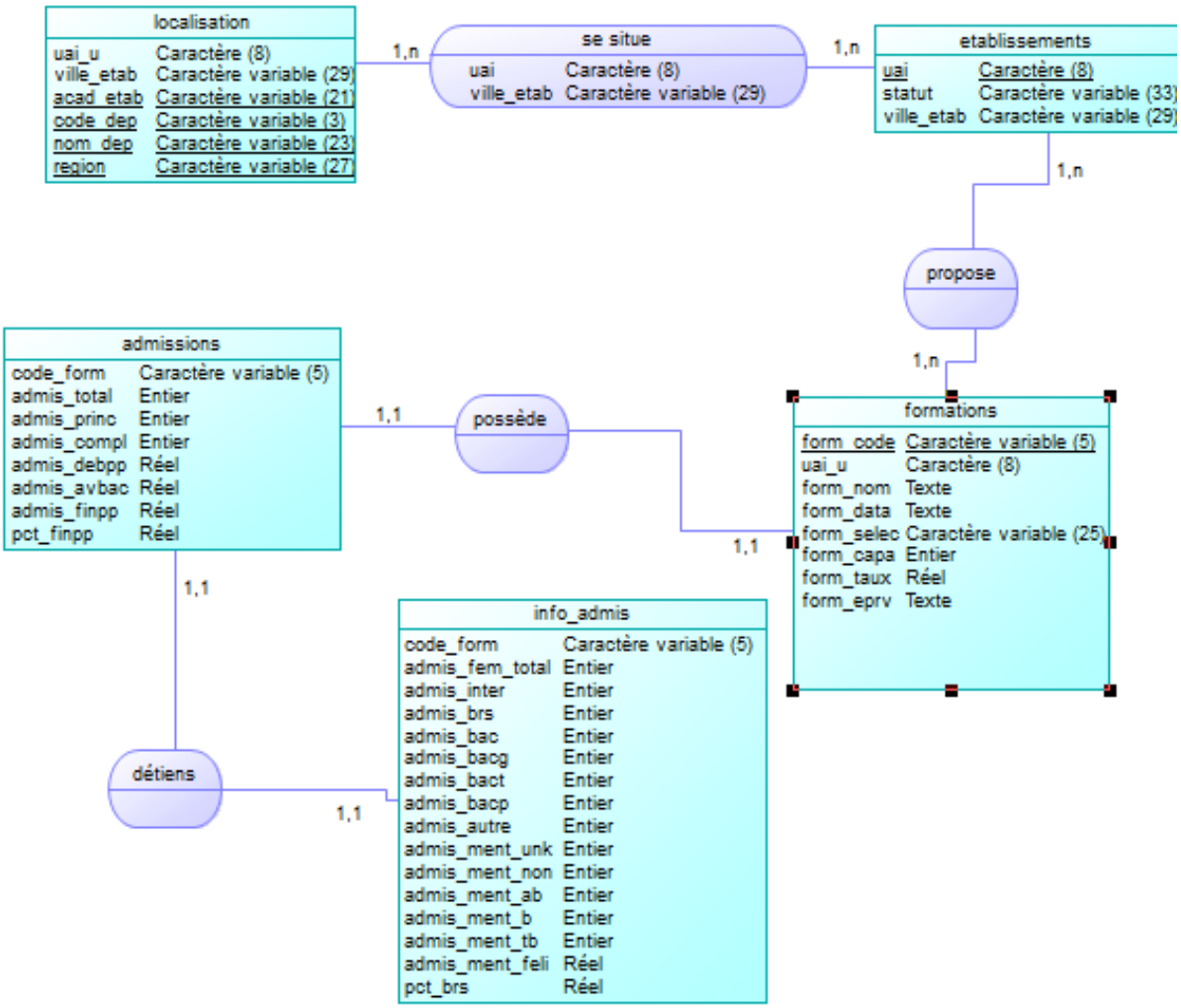
SAé S2.04 : Exploitation de BDD

Philippe Mathieu
2023–2024

Lucas De Jesus Teixeira

Groupe D

Louis Beck



Exercice 1 : Comprendre les données

Question 1 :

1. Combien y-a t-il de lignes ? Justifiez !

Il y a 13870 lignes et la commande qu'on a utilisé était `wc -l fr-esr-parcoursup.csv`.

2. Que représente une ligne ?

Chaque ligne représente une formation ainsi que l'établissement et les caractéristiques auxquelles elles appartiennent.

3. Combien y-a t-il de colonnes ? Justifiez !

Il y a 118 colonnes, on utilise la commande suivante :

`head -1 fr-esr-parcoursup.csv | tr ';' '\n' | wc -l`

4. Quelle colonne identifie un établissement ? (numéro et nom de col)

La 3ème colonne identifie l'établissement (*Code UAI de l'établissement*)

5. Quelle colonne identifie une formation ? (numéro et nom de col)

La 110ème colonne identifie la formation (*cod_aff_form*)

6. Combien de lignes font référence à notre BUT Informatique ?

1 seule ligne fait référence à notre BUT Informatique

On utilisait la commande suivante sur la table import :

```

select count(*) from import where n4

like '%Institut universitaire de technologie de Lille%'
```

AND n10 like '%BUT%Info%';

```

**7. Quelle colonne identifie un département ? (numéro et nom)**

La 5ème colonne permet d'identifier le département (Code départementale de l'établissement)

**8. Comment envisagez vous importer ces données ?**

En utilisant la commande \copy de PSQL et en récupérant le fichier avec wget

**9. Quels problèmes identifiez vous dans ces données initiales ? (il y en a surement plusieurs, expliquez les clairement)**

Les pourcentages sont des symboles complexes donc au lieu de mettre un nombre à pourcentage, on met un entier. Il faut aussi vérifier et corriger en cas d'erreur de données de saisies et formatages et vérifier la cohérences des données.

---

### Question 2 :

**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.**

On a utilisé la commande suivante pour ne garder que la 1ère ligne avec les noms de colonnes :

`head -n 1 fr-esr-parcoursup.csv >dico.xls`

Puis on est rentré dans le fichier et on a nommé la colonne A2 par **n1** et on a étiré jusqu'à arrivé à la dernière l'occution dans la 1ère ligne (DN).

**2. Créer une table import permettant l’importation de ces données (fournir le code)**
```

CREATE temp TABLE import (

n1 text, n2 text, n3 text, n4 text, n5 text, n6 text, n7 text, n8 text, n9 text,

n10 text, n11 text, n12 text, n13 text, n14 text, n15 text, n16 text, n17 text,

n18 text, n19 text, n20 text, n21 text, n22 text, n23 text, n24 text, n25 text,

n26 text, n27 text, n28 text, n29 text, n30 text, n31 text, n32 text, n33 text,

n34 text, n35 text, n36 text, n37 text, n38 text, n39 text, n40 text, n41 text,

n42 text, n43 text, n44 text, n45 text, n46 text, n47 text, n48 text, n49 text,

n50 text, n51 text, n52 text, n53 text, n54 text, n55 text, n56 text,

n57 text, n58 text, n59 text, n60 text, n61 text, n62 text, n63 text,

n64 text, n65 text, n66 text, n67 text, n68 text, n69 text, n70 text,

n71 text, n72 text, n73 text, n74 text, n75 text, n76 text, n77 text,

n78 text, n79 text, n80 text, n81 text, n82 text, n83 text, n84 text,

n85 text, n86 text, n87 text, n88 text, n89 text, n90 text, n91 text,

n92 text, n93 text, n94 text, n95 text, n96 text, n97 text, n98 text,

n99 text, n100 text, n101 text, n102 text, n103 text, n104 text, n105 text,

n106 text, n107 text, n108 text, n109 text, n110 text, n111 text, n112 text,

n113 text, n114 text, n115 text, n116 text, n117 text, n118 text);

On a ensuite modifié les « text » avec des types plus adaptés (voir prochaine question).

3. S’assurer que les types de colonnes soient les plus restrictifs possibles

CREATE temp TABLE import (

n1 INT, n2 VARCHAR(33), n3 CHAR(8), n4 TEXT, n5 VARCHAR(3), n6 VARCHAR(23), n7 VARCHAR(27), n8 VARCHAR(21), n9 VARCHAR(29),

n10 TEXT, n11 VARCHAR(25), n12 VARCHAR(18), n13 TEXT, n14 TEXT, n15 TEXT, n16 TEXT, n17 VARCHAR(21),

n18 INT, n19 INT, n20 INT, n21 INT, n22 TEXT, n23 INT, n24 INT, n25 INT,

n26 INT, n27 INT, n28 INT, n29 INT, n30 INT, n31 INT, n32 INT, n33 INT,

n34 INT, n35 INT, n36 INT, n37 TEXT, n38 TEXT, n39 INT, n40 INT, n41 INT,

n42 INT, n43 INT, n44 INT, n45 INT, n46 INT, n47 INT, n48 INT, n49 INT,

n50 INT, n51 FLOAT, n52 FLOAT, n53 FLOAT, n54 TEXT, n55 INT, n56 INT,

n57 INT, n58 INT, n59 INT, n60 INT, n61 INT, n62 INT, n63 INT,

n64 INT, n65 INT, n66 FLOAT, n67 INT, n68 INT, n69 INT, n70 TEXT,

n71 TEXT, n72 INT, n73 INT, n74 FLOAT, n75 FLOAT, n76 FLOAT, n77 FLOAT,

n78 FLOAT, n79 FLOAT, n80 FLOAT, n81 FLOAT, n82 FLOAT, n83 FLOAT, n84 FLOAT,

n85 FLOAT, n86 FLOAT, n87 FLOAT, n88 FLOAT, n89 FLOAT, n90 FLOAT, n91 FLOAT,

n92 FLOAT, n93 FLOAT, n94 FLOAT, n95 FLOAT, n96 FLOAT, n97 FLOAT, n98 FLOAT,

n99 FLOAT, n100 FLOAT, n101 FLOAT, n102 VARCHAR(39), n103 TEXT, n104 TEXT, n105 TEXT,

n106 TEXT, n107 TEXT, n108 VARCHAR(45), n109 VARCHAR(19), n110 CHAR(5), n111 TEXT, n112 TEXT,

n113 TEXT, n114 FLOAT, n115 FLOAT, n116 FLOAT, n117 CHAR(5), n118 CHAR(5));

4. Remplir cette table avec les données récupérées (fournir le code)

`\copy import from fr-esr-parcoursup.csv with(FORMAT CSV, NULL 'null_string', delimiter ';', HEADER)`

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 ?

```
SELECT COUNT(DISTINCT n10)

FROM import;
```

Il y a 3207 formations gérés par ParcourSup.

(b) Combien il y a d’établissements gérés par ParcourSup ?

```
SELECT COUNT(DISTINCT n4)
FROM import;
```

Il y a 3602 établissements gérés par ParcourSup.

(c) Combien il y a de formations pour l’université de Lille ?

```
SELECT COUNT(DISTINCT n10)
FROM import
WHERE n4 LIKE '%Univ% de Lille%';
```

Il y a 124 formations pour l'Université de Lille.

(d) Combien il y a de formations pour notre IUT ?

```
SELECT COUNT(DISTINCT n10)
FROM import
WHERE n4 LIKE 'Institut universitaire de technologie de Lille%';
```

Il y a 10 formations pour notre IUT.

(e) Quel est le code du BUT Informatique de l’unversité de Lille ?

```
SELECT n110
FROM import
```

WHERE n13 LIKE 'Institut universitaire de technologie de Lille%BUT%Info%';

```

Le code du BUT Informatique de l'université de Lille est 6888.

(f) Citez 5 colonnes contenant des valeurs nulles

Pour chaque colonne "Text" ou "Varchar", on utilise la commande sql :

```
SELECT n9 FROM import ORDER BY length(n9) LIMIT 1;
```

Si la valeur donnée est vide, alors la table contient une valeur nulle :

n9, n16, n17, n54, n70

---

## Exercice 2 : Ventiler les données

### Question 2 :

1. Quelle taille en octet fait le fichier récupéré ?

12 423 586 octets

2. Quelle taille en octet fait la table import ?

16 408 576 octets

3. Quelle taille en octet fait la somme des tables créées ?

7 372 800 octets

4. Quelle taille en octet fait la somme des tailles des fichiers exportés correspondant à ces tables ?

4 296 373 octets
```

Exercice 3 : Requêtage

(Voir requetes.sql)