IUT 2 Grenoble Université Pierre-Mendès-France

Année 2018/2019

Exercice
Transformation des données

Exercice 1

Vous disposez d'extraits de 2 fichiers[1] de données : customer.cvs et state.txt. Le premier fichier comporte des données concernant des clients dont un numéro représentant un identifiant de l'état des Etats-Unis du client ; le deuxième fait le lien entre les identifiants des états américains et leur nom.

Customer.cvs

```
/**************************************************************/;;;;;;;;
/*******  Extract on Mon Oct 02 10:30:19 CEST 2006  ***************/;;;;;;;;
/**************************************************************/;;;;;;;;

id;CustomerName;CustomerAddress;idState;id2;RegTime;RegisterTime;Sum1;Sum2
1;Griffith Paving and Sealcoatin;talend@apres91;7;41;03/11/2006 09:20;2001-01-17 06:26:40.000;67852;61521.4852
2;Bill's Dive Shop;511 Maple Ave.  Apt. 1B;35;5;19/11/2004 15:48;2002-06-07 09:40:00.000;88792;15434.1000
3;Childress Child Day Care;null;1;28;16/02/2005 08:27;1990-04-01 21:00:00.000;35340;17856.8818
4;Facelift Kitchen and Bath;unknown;0;15;22/08/2002 09:55;1972-04-23 18:00:00.000;6097.;55560.2387
5;Terrinni & Son Auto and Truck;770 Exmoor Rd.;0;9;28/06/2001 09:15;1982-04-19 10:26:40.000;5146.;39098.1148
6;Kermit the Pet Shop;1860 Parkside Ln.;28;15;17/08/2003 10:07;2006-05-27 17:00:00.000;16087;29924.9294
7;Tub's Furniture Store;807 Old Trail Rd.;15;9;27/08/2000 03:13;1970-03-27 23:08:16.000;53216;65352.5674
8;Toggle & Myerson Ltd;null.;9;15;24/03/2006 23:07;2005-08-02 01:26:40.000;74168;77920.6026
9;Childress Child Day Care
10;Elle Hypnosis and Therapy Cent;2032 Northbrook Ct.;1;7;11/01/1977 03:07;1975-06-10 20:20:00.000;48498;45844.9148
11;Lennox Air Pollution Control;4522 N. Greenview Apt. 1B;48;35;20/07/1987 07:13;1983-02-26 17:08:16.000;23992;93520.1160
12;Keyth Contracting and Repair;1547 Knolwood Rd.;25;39;12/01/2000 15:33;2001-09-10 11:01:36.000;27786;46530.0991
13;Park District Of America;2678 Sheridan Rd.;46;41;08/06/2003 11:15;2005-10-02 00:34:56.000;5448.;29062.8166
14;Nirabi Auto Service;1915 Lewis Ln.  Apt 13;8;5;19/06/2002 07:15;1997-02-18 21:06:40.000;23220;43086.6464
15;Darcy Frame and Matting Servic;1633 McGovern place;21;28;20/10/1987 12:22;2001-01-07 20:40:00.000;46096;31014.8296
16;Glenwood Credit Union;511 Maple Ave.  Apt. 1B;46;15;17/07/2000 17:27;2003-12-03 19:08:16.000;90440;31684.7929
17;Gourmet the Frog;788 Tennyson Ave.;1;9;07/09/2002 03:55;1983-01-31 22:26:40.000;67680;62038.7700
18;Acturial Enterprises Ltd.;3385 University Ave.;34;12;07/04/2001 10:42;2004-05-31 15:00:00.000;45292;17008.7659
```

state.txt

```
idState;LabelState
1;Alabama
2;Alaska
3;Arizona
4;Arkansas
5;California
6;Colorado
7;Connecticut
8;Delaware
```

[1] Issus des tutoriels de Talend : http://www.talendforge.org/tutorials/menu.php

9;Florida
10;Georgia
11;Hawaii
12;Idaho
13;Illinois
14;Indiana
15;Iowa
16;Kansas
17;Kentucky
18;Louisiana
19;Maine
20;Maryland

A partir des données contenues dans ces 2 fichiers, l'objectif est d'obtenir un nouveau fichier *ClientFloride.cvs* contenant :

- Les identifiants des clients de Floride
- leur nom
- leur adresse
- le libellé de l'état dans lequel ils habitent

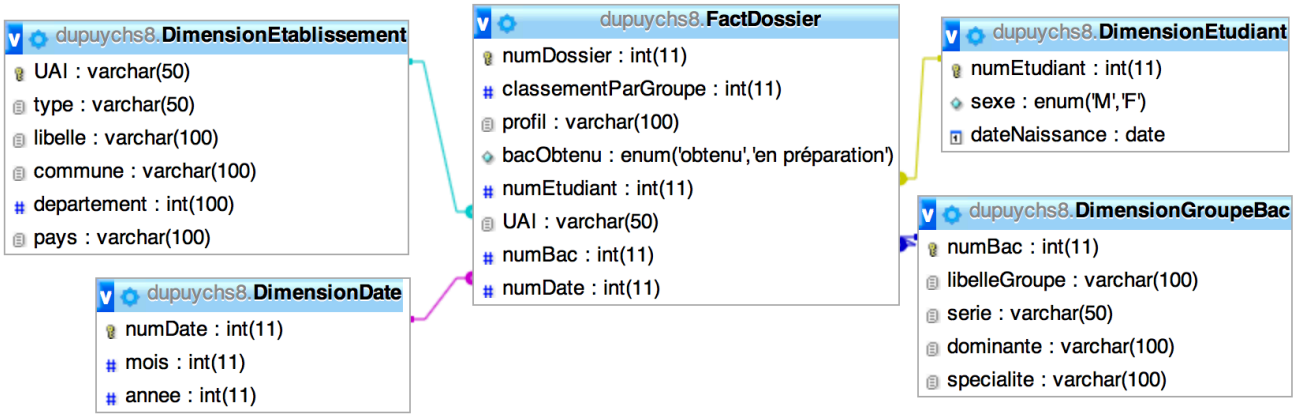
Questions

1. *Quels sont les délimiteurs de données ?*
2. *Définissez la matrice de transformation pour obtenir le nouveau fichier ClientFloride.cvs*
3. *Quelles sont les données insérées dans le fichier cible ?*

Exercice 2

Pour leur formation après le Baccalauréat, les futurs étudiants remplissent leurs vœux au sein de l'application Post-Bac. Le département informatique a lui aussi une vision de Post-Bac qui lui permet de sélectionner ses candidats. Il a les informations sur le candidat, son établissement d'origine. Les candidats sélectionnés sont classés suivant leur Bac d'origine. Il peut ensuite extraire un fichier avec les données des candidats. Un tel fichier anonymisé vous est fourni sous Chamilo sous le nom 2015-09-Classes-REcus.xls

Les données issues de ces fichiers doivent être chargées dans un entrepôt de données dont un extrait du schéma est fourni ci-dessous. Le numéro de dossier est un nombre auto-incrémenté fourni par le système de gestion de données.



Le chargement des données dans les tables DimensionEtablissement, DimensionEtudiant, DimensionGroupeBac ne demande aucune transformation.

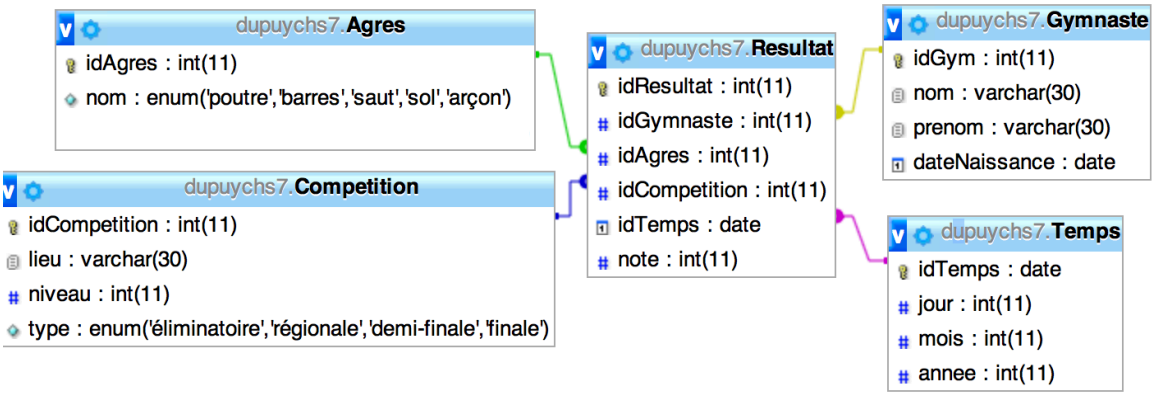
Question :
Définissez les matrices de transformations pour obtenir les tables :

- *DimensionDate*
- *FactDossier*

Exercice 3

Un club de gymnastique souhaite améliorer ses résultats sportifs en analysant les résultats de ses gymnastes et de ses équipes aux compétitions. Les compétitions regroupent des gymnastes qui sont jugées par niveau. Elles peuvent avoir lieu sur plusieurs jours. Lors d'une compétition, chaque gymnaste est évaluée sur des agrès : par exemple, les filles sont évaluées au sol, au saut de cheval, à la poutre et aux barres asymétriques. Il obtient une note individuelle pour chacune de ses épreuves.

Pour analyser les résultats sportifs, le club a mis en place un entrepôt sous MySQL dont le schéma en étoile est le suivant. Les données correspondant aux gymnastes, aux agrès et aux compétitions sont enregistrées avant le début des compétitions.



Attention : tous les identifiants (sauf idTemps) sont des nombres auto-incrémentés créés le système de gestion de base de données.

Lors d'une compétition, un gymnaste est évalué au fur et à mesure et ses résultats aux différents agrès sont notés sur une feuille. A la fin de la compétition, la feuille de chaque gymnaste est donnée au club afin que l'entraineur saisisse les résultats dans un fichier Excel nommé Resultats.xsl. On suppose que les noms et les prénoms sont saisis correctement et que la date de la compétition est au format Date.

Résultats

Compétition	Date	Gymnaste	Agrès	Note
3	12/3/14	Léa Flaubi	Poutre	6.3
3	13/3/14	Léa Flaubi	sol	7.1
3	13/3/14	Léa Flaubi	barres	7.3
3	12/3/14	Léa Flaubi	Saut	6.8
3	13/3/14	Nathalie Gie	SOL	7.3
3	13/3/14	Nathalie Gie	Barres	6.9

Les tables Agres, Competition, Gymnaste et Temps ont été remplies précédemment.

Question : Décrivez la matrice de transformation qui permet de remplir la table Resultat à l'issue des compétitions.