Project Data Warehouse

Fabien Tang, M1 BI2

The main goal of this DW is to better structure the data to optimize and make future querries easier. We'll also use later the K-Mean and K-NN algorithm to determine different profile of blood donor and which kind of profile was giving on March 2007.

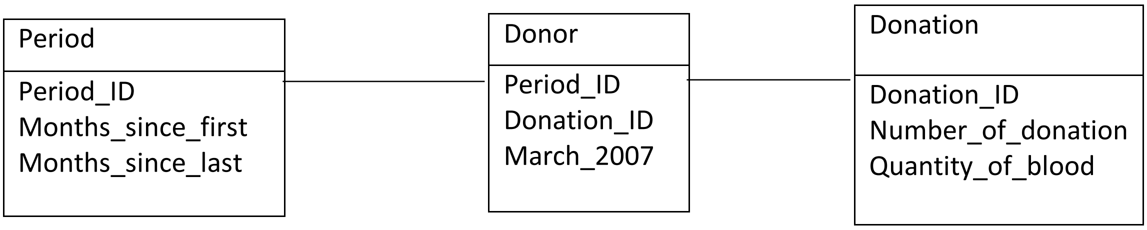
1. Transfusion.data

Structuration

Recency	Frequency	Monetary	Time	A
2	50	12500	98	1

Recency : Months since last donation
Frequency : Total number of donation
Monetary : Total blood donated in c.c
Time : Months since first donation
A : Whether he/she donated blood in March 2007

2. Logical design of the database



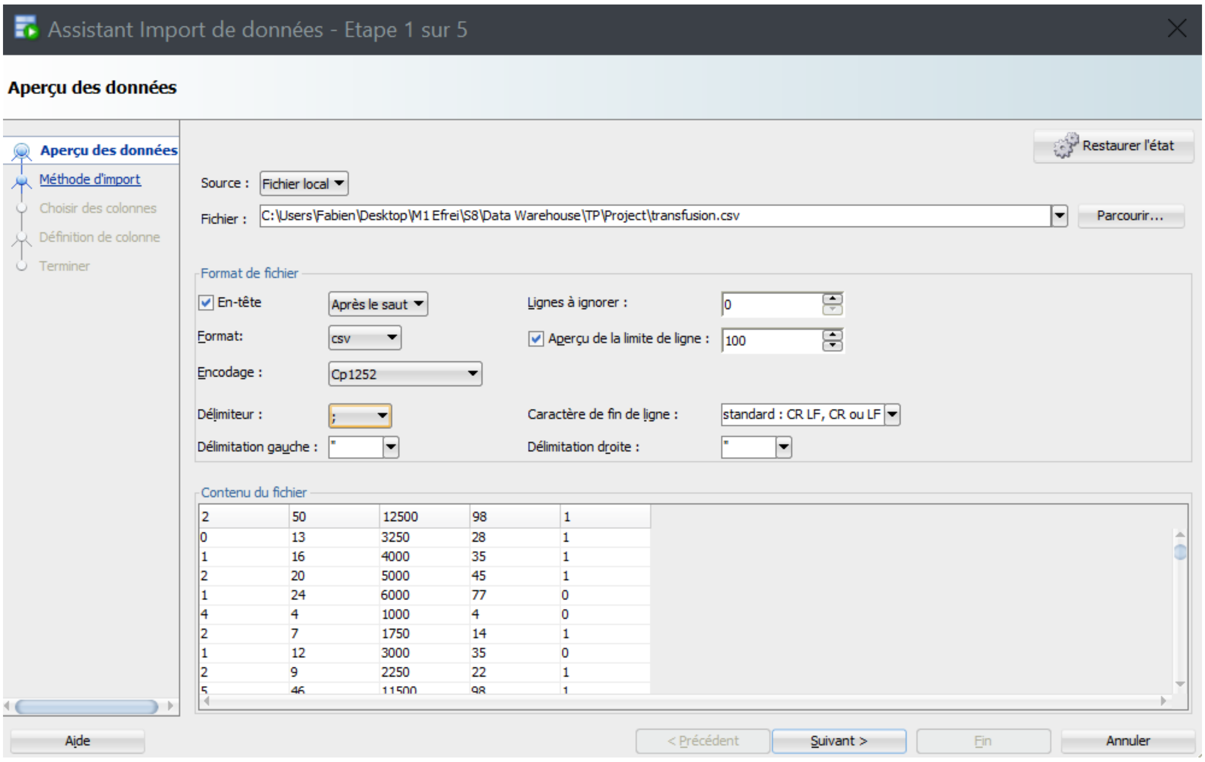
3. Creation of the Data Warehouse (create_table.sql)

```
CREATE TABLE Period
(
        Period_ID number(7) primary key,
        Months_since_first number(7),
        Months_since_last number(7)
) ;

CREATE TABLE Donation
(
        Donation_ID number (7) primary key,
        Number_of_donation number(7),
        Quantity_of_blood number(7)
) ;

CREATE TABLE Donor
(
        Period_ID number(7),
        Donation_ID number(7),
        March_2007 number(1),
        foreign key (Period_ID) references Period(Period_ID),
        foreign key (Donation_ID) references Donation(Donation_ID)
) ;
```

4. Insert data into the Data Warehouse (insert_table.sql)

<u>To insert the data into the DW :</u>

- Convert Transfusion.data into Transfusion.txt file (just change the file extension)
- Open the Transfusion.txt with Excel then save it as a Transfusion.csv
- From SQLDevelopper import Transfusion.csv into a tempory table to store all the data (Photo 1)
- Create new attributes Period_ID/Donation_ID auto-incremented (Photo 2)
- Insert into the others tables the data from the tempory table
- Export the tables to have a backup (**export.sql**)



Assistant Import de données - Etape 1 sur 5

Aperçu des données

Source : Fichier local

Fichier : C:\Users\Fabien\Desktop\M1 Efrei\S8\Data Warehouse\TP\Project\transfusion.csv

Format de fichier

☑ En-tête: Après le saut

Lignes à ignorer : 0

Format: csv

☑ Aperçu de la limite de ligne : 100

Encodage : Cp1252

Délimiteur : ;

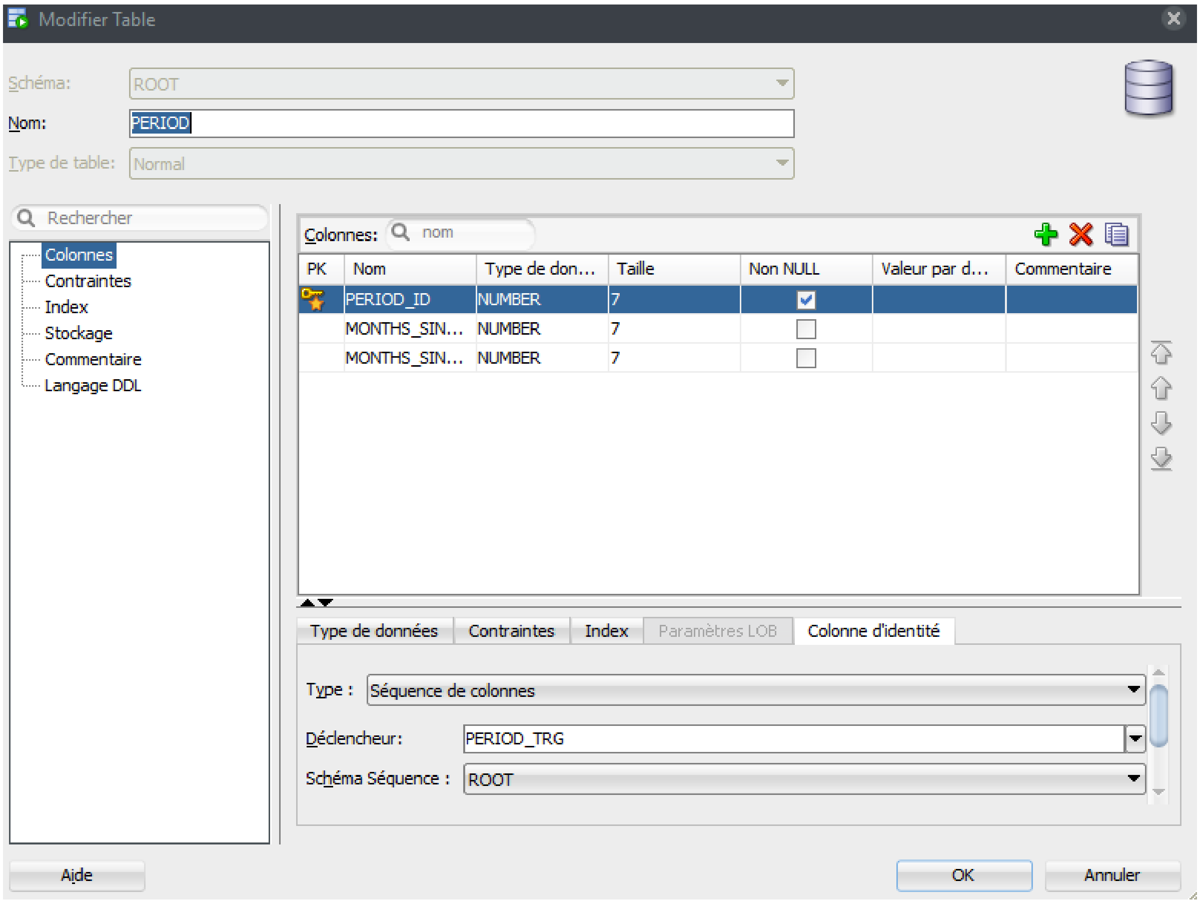
Caractère de fin de ligne : standard : CR LF, CR ou LF

Délimitation gauche : "

Délimitation droite : "

Contenu du fichier

2	50	12500	98	1
0	13	3250	28	1
1	16	4000	35	1
2	20	5000	45	1
1	24	6000	77	0
4	4	1000	4	0
2	7	1750	14	1
1	12	3000	35	0
2	9	2250	22	1
5	46	11500	98	1



Modifier Table

Schéma: ROOT

Nom: PERIOD

Type de table: Normal

Colonnes: nom

PK	Nom	Type de don...	Taille	Non NULL	Valeur par d...	Commentaire
🔑	PERIOD_ID	NUMBER	7	☑		
	MONTHS_SIN...	NUMBER	7	□		
	MONTHS_SIN...	NUMBER	7	□		

Type : Séquence de colonnes

Déclencheur: PERIOD_TRG

Schéma Séquence : ROOT

```
Insert into Period (months_since_first, months_since_last)
Select time,recency
From transfusion;

Insert into Donation (number_of_donation, quantity_of_blood)
Select frequency, monetary
From transfusion;

Create sequence seq_fact start with 1;
Create sequence seq_fact2 start with 1;

Insert into Donor (period_id, donation_id, march_2007)
Select seq_fact.nextval, seq_fact2.nextval, t.march
From transfusion t;
```

5. **Query the Data Warehouse using OLAP (CUBE,…) queries in order to obtain in output a rectangular matrix X with n lines and p columns containing in lines the objects and in columns theirs characteristics. These queries depends on the data problem and application field. You can obtain several matrices. Explain the queries and the results.**

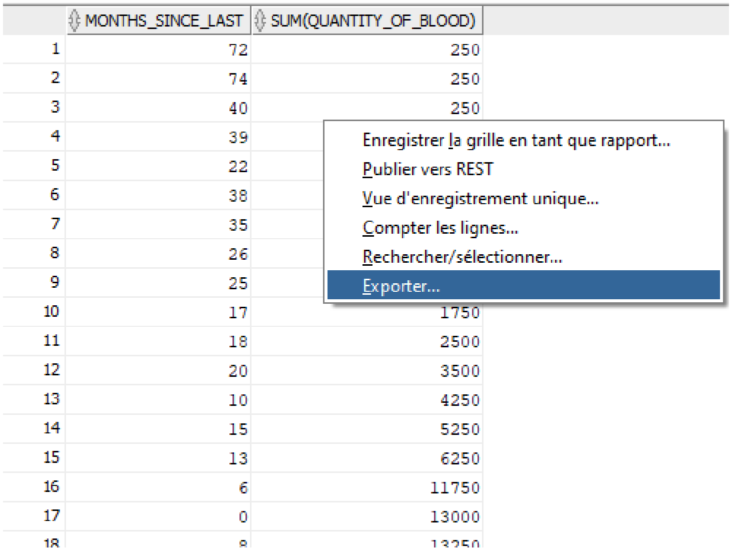
The aim with the first matrix will be with the K-Means method to determine different profil of donator based on the months since last donation and the total amount of blood given.

```
Select months_since_last, sum(quantity_of_blood)
From period, donor, donation
Where donor.period_id=period.period_id and donation.donation_id=donor.donation_id
Group by cube(months_since_last)
Order by 2;
```

The second matrix, will be usefull to determine with K-NN Method which kind of profile based on months since last donation and total number of donation is giving on March_2007.

```
Select pe.months_since_last, dona.number_of_donation, dono.march_2007
From period pe, donation dona, donor dono
Where pe.period_id=dono.period_id and dono.donation_id=dona.donation_id
Order by 2;
```

Export through sql developper : (matrix 1 : analyse_1.csv, matrix 2 : analyse_2.csv)



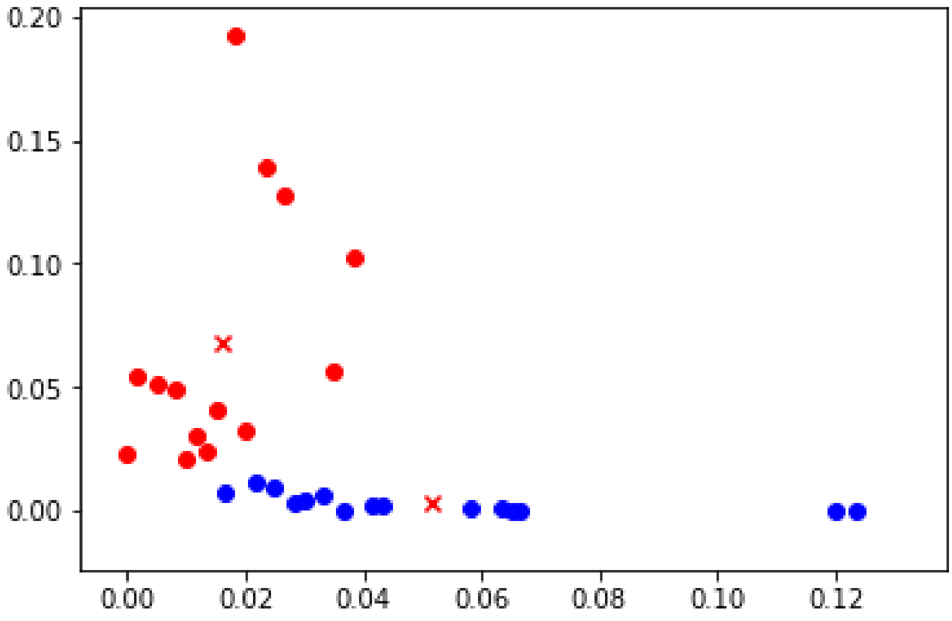
	MONTHS_SINCE_LAST	SUM(QUANTITY_OF_BLOOD)
1	72	250
2	74	250
3	40	250
4	39	
5	22	
6	38	
7	35	
8	26	
9	25	
10	17	1750
11	18	2500
12	20	3500
13	10	4250
14	15	5250
15	13	6250
16	6	11750
17	0	13000
18	8	13250

Enregistrer la grille en tant que rapport...
Publier vers REST
Vue d'enregistrement unique...
Compter les lignes...
Rechercher/sélectionner...
Exporter...

6. **Use Phyton in order to load the obtained matrix (matrices) and visualize the data using scatter plot. Analyse the results. Apply a clustering model using k-means method and visualize the results. Explain the results.**

Analysis 1 – K-MEAN (Python : Analyse_1.ipynb)

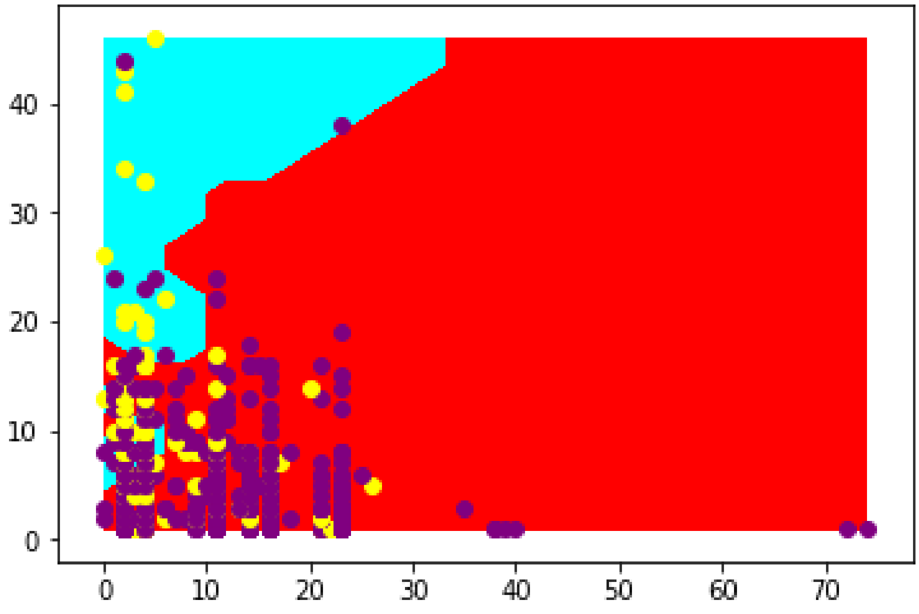
In this first analysis the K-Mean method gave us those 2 profiles of populations (2 clusters)



X = Months since last donation (x600) Y = Sum (quantity of blood) (x550250) (multiply for denormalized form)

Analysis 2 – K-NN (Python : Analyse_2.ipynb)

In the second analysis with the K-NN Method we can easily say that if the number of donation is high and month since last donation low, the people would gave their blood on March 2007.



X = Months since last donation / Y = number of donation

Yellow dot = gave his blood on March 2007 / Purple dot = did not gave his blood on March 2007

Blue part : prediction the person will give his blood // Red part : prediction the person will not give his blood

Conclusion

In conclusion, through this project we sucessfully structured the transfusion database with a Star Schema modelization. We then used OLAP Query through group by and Cube to produce csv that were used to do some relevant analysis. Finally, the analysis shown us 2 major profile of population and help us to easily see which profile gave their blood on March 2007