

PCA : FOUILLE DES BIOMARQUEURS

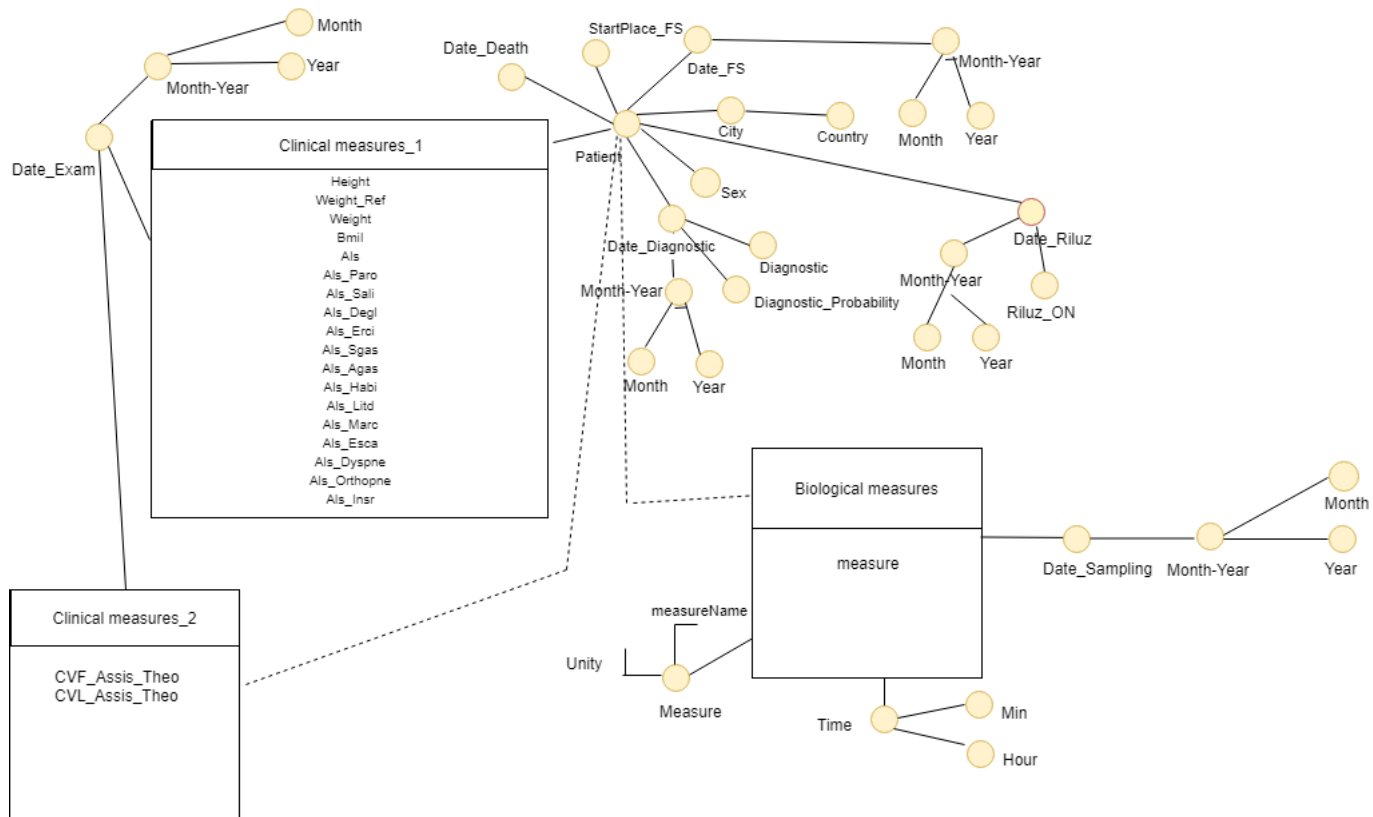
Contents

1	Présentation du modèle conceptuel de l'entrepôt des données	2
2	Dictionnaire de données	2
3	Intégration	4
3.1	Dénormalisation des fichiers des données	4
3.2	Création de schéma de la base de données	6
3.3	Transformation et chargement des données	6
3.3.1	Transformation effectués:	6
3.3.2	Execution de Jobs avec la ligne de commande:	6
3.4	Accès aux données:	7

List of Figures

1	le modèle conceptuel de l'entrepôt de données	2
2	Le structure de premier fichier avant et après Transformation	5

1 Présentation du modèle conceptuel de l'entrepôt des données



FS : First Symptom

Figure 1: le modèle conceptuel de l'entrepôt de données

2 Dictionnaire de données

Nom	Description	Dimension	Mesure	Attribut
Date_Exam	la date d'un examen clinique pour un patient donné	OUI	NON	NON
Date_Riluz	la première date quand le patient a commencé de prendre le riluzole	OUI	NON	NON

Nom	Description	Dimension	Mesure	Attribut
Date_Diagnostic	La date de diagnostic de la maladie ALS	NON	NON	OUI
Patient	Une dimension qui contient les informations sur un patient (sexe,date de naissance etc)	OUI	NON	NON
StartPlace_FS	Le lieu de début de la maladie ALS dans le corps du patient	NON	NON	OUI
Date_FS	la date de début des premiers symptômes	NON	NON	OUI
Diagnostic	le diagnostic de la maladie (Exemple: ALS)	NON	NON	OUI
Diagnostic_Probability	La probabilité de diagnostic (Exemple: probable)	NON	NON	OUI
Measure	Une dimension qui contient les mesures biologiques (Exemple :Ferritinine) et leurs unités	OUI	NON	NON
Time	le temps de prélèvement (Exemple : 10h30)	OUI	NON	NON
Date_Sampling	la date de prélèvement	OUI	NON	NON
measure	La valeur d'une mesure biologique donnée	NON	OUI	NON
Riluz_ON	Cet attribut prend "OUI" dans le cas ou le patient prend le riluzole , "NON" dans le cas contraire	NON	NON	OUI
Height	l'hauteur du patient à une date donnée	NON	OUI	NON
Weight_Ref	l'hauteur de référence pris 6 mois avant une date d'examen donnée	NON	OUI	NON
BMI	l'indice de masse corporelle du patient à une date donnée	NON	OUI	NON
Als	le score ALS du patient à une date donnée calculé à partir les détails Als.	NON	OUI	NON
Cvl_Assis_Theo	C'est la capacité Vitale lente en pourcentage.	NON	OUI	NON
Cvf_Assis_Theo	c'est la capacité Vitale Forcée en pourcentage.	NON	OUI	NON
Date_Death	la date de décès de Patient	NON	OUI	NON

Nom	Description	Dimension	Mesure	Attribut
Als_Paro...Als_Insr	13 items qui représentent les détails du score Als,chaque item est un score de 0 à 4 . La somme donne le score Als.	NON	OUI	NON

Table 1: Dictionnaire des données de l’entrepôt construit.

3 Intégration

3.1 Dénormalisation des fichiers des données

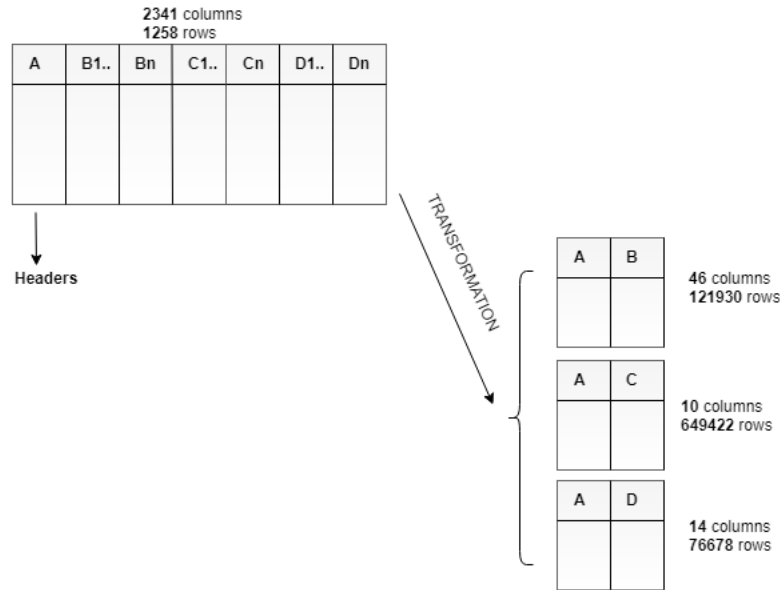
Langage: Python.

Fichiers reçus de l’hôpital:

- Un fichier contenant les patients avec leurs mesures cliniques à des dates d’examens différentes.
- Un fichier complémentaire contenant les dates de décès des patients.

Dans ce qui suit,nous allons montrer les structure des deux fichiers avant et après toute transformation effectuée à l’aide d’un script python implementée pour cette tâche.

Structure du premier fichier: Dans la figure ci-dessous,la matrice à gauche montre la structure initiale du fichier reçu. Chaque groupe de colonne :A,B,C où D est décrit en détail en bas de la figure. Les trois matrices à droite montre la structure des fichiers résultats qui servent à des inputs pour nos jobs Talend.



Columns Details :

A	B	C	D
<ul style="list-style-type: none"> • PATIENT • REF_NAME • REF_ID770S4V3638 • REF_SURNAME • NAME • SURNAME • SEX • DOB • NO • IPP • AUTRDIAG • ND_COMMENT • DIAGPROBA 	<ul style="list-style-type: none"> • DATEXAM • HEIGHT • WEIGHT_REF • WEIGHT • BMI • ALS • ALS DETAILS 	<ul style="list-style-type: none"> • DATE_RILUZ 	<ul style="list-style-type: none"> • DATE_PREVENT • CVL_ASSIS_THEO • CVF_ASSIS_THEO(FORCED VITAL CAPACITY)

Figure 2: Le structure de premier fichier avant et après Transformation

Structure de deuxième fichier: Le deuxième fichier contient seulement trois groupes de colonnes ,le dernier groupe est la date de décès. Nous avons suivi la meme démarche que le fichier précédent.

Etapes incluses dans la transformation des fichiers

- Encodage des patients: Nous avons utilisé une fonction de hashage md5 appliquée sur les IPPs des patients. Pour les patients avec un IPP null, la clé est appliquée sur le nom+prénom+date de naissance du patient. Toutes les données qu'on utilise sont anonymisées suivant cette fonction de hashage.

-
- génération des fichiers: Les transformations effectuées ensuite à l'aide du script python sur des données déjà anonymisées donnent en sortie les fichiers montrés dans la figure 2. Ces fichiers sont l'entrée de notre pipeline d'intégration créée avec Talend.

3.2 Création de schéma de la base de données

Outil : Sql Power Architect est un outil qui permet la création de schémas des entrepôts de données, d'appliquer l'ingénierie avancée sur les schémas créés, comparer deux schémas de bases de données etc ?

Cet outil était utilisée pour la création et la manipulation du schéma conçu.

3.3 Transformation et chargement des données

Outil : Talend.

3.3.1 Transformation effectués:

- Des dates qui respectent pas le format dd/mm/yyyy .On trouve comme exemple de donnée la date ND/02/2011. Ce type de valeurs est remplacée par ND ,mais on garde l'année et mois dans deux colonnes à part.
- transformation de valeurs de chaîne de caractères en valeurs numériques (Exemple : le poids).
- La gestion des duplications: Des patients qui se trouvent dans plusieurs lignes avec la même date d'examens ; ces patients sont gardés à part dans un fichier pour plus d'investigation après.
- La gestion des valeurs nulls : .Les valeurs nulls sont remplacées par "ND" ; "Not Defined". On n'a pas gardé les mesures cliniques des patients dont la date d'examen n'est pas renseignée.

3.3.2 Execution de Jobs avec la ligne de commande:

Pour exécuter un job avec la ligne de commande, Il faut :

- Stocker les paramètres de connexion à la base ou aux fichiers dans des variables de contextes.

-
- exporter le jobs en jar.
 - executer le job avec la commande " java -c Nom-de-job "

Si vous souhaitez modifier l'emplacement d'un fichier de données ou les paramètres de connexion ,il suffit de modifier le fichier "properties" exporté avec le jar.

3.4 Accès aux données:

- Installer un client vpn,exemple: OpenVPN.
- Se connecter avec les identifiants de l'université.
- installer x2GO puis se connecter avec les identifiants de l'université.