

# PCA : FOUILLE DES BIOMARQUEURS

---

## Contents

<b>1</b>	<b>Présentation du modèle conceptuel de l'entrepôt des données</b>	<b>2</b>
<b>2</b>	<b>Dictionnaire de données</b>	<b>2</b>
<b>3</b>	<b>Intégration</b>	<b>4</b>
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

## List of Figures

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

FS : First Symptom

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	une mesure clinique de patient à une date donnée	NON	OUI	NON
Cvf_Assis_Theo	une mesure clinique de patient à une date donnée	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.



---

## 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 meme date d'examens ; ces patients sont gardés à part dans un fichier pour plus d'investigation après.
- La gestion des valeurs nulls : pour les valeurs numériques, on a gardé les valeurs nulls pour faciliter les calculs après.Les valeurs de chaîne de caractères 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 executer un job avec la ligne de commande, Il faut :

- Stoker 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.