Université Gaston Berger de Saint-Louis

UFR Sciences Appliquées et Technologie

Section Informatique

Master GDIL 2025

Base de données multidimensionnelle

TP 1 : *Extract-Transform-Load (ETL avec Pentaho for Data Integration (PDI))*

Exercice 0 :

Lancer ***Pentaho for Data Integration (PDI)*** et essayer de vous familiariser à l'environnement de ***Pentaho*** (avec l'aide du professeur).

Exercice 1

Considérons maintenant les données sur les prix des drogues aux USA en fonction de l'état, de la date et de la qualité du produit (**weed_price.csv** fourni en annexe). Nous désirons charger ces données dans un Entrepôt de données à des fins d'analyse.

1. Ecrire un Job permettant de remplacer les valeurs manquantes **« NA »** de la colonne **_ LowQ_** par 0 *(on pourra utiliser **Replace in String**).*
2. Ecrire un nouveau Job permettant de supprimer les guillemets se trouvant au niveau de la colonne **_STATE_** *(on pourra utiliser **Replace in String** également).*
3. En utilisant **Pentaho** écrire un programme permettant de charger le fichier de données dans une base de données de votre choix.
4. Charger dans une table multidimensionnelle **(resume)** les 10 États ayant enregistré le plus grand nombre total d'observations de ventes de cannabis (toutes qualités confondues) sous la forme suivantes :

État	HighQN	MedQN	LowQN	Total des ventes observées
California	6711236	7529650	438358	**14679244**
Florida	3778349	3200120	283803	7262272
Texas	3354655	3388923	472326	7215904
New York	3189037	3431299	255090	6875426
Illinois	2267183	2158738	184243	4610164
Pennsylvania	2209594	1943371	205207	4358172
Georgia	1784813	1360690	124824	3270327
Ohio	1717201	1690227	222286	3629714
Michigan	1644780	1475057	121569	3241406
North Carolina	1628101	1175056	126621	2929778

Exercice 2

1. Prise en main de **PDI**.
2. Créer un nouveau Job permettant de trier le fichier Client en utilisant : **Tri Lignes** (Tri sur le nom)
3. Créer un Job qui permet d'effectuer la jointure des données contenues dans les fichiers Client et Etat en utilisant : (**Jointure Multiples**)
4. Filtrer les données de jointure obtenues en enlevant les clients ayant la valeur ***nomEtat*** non renseignée (Puis on affichera les données rejetés (***NomClient*** *et* ***NomEtat***) à l'écran en utilisant ***Contrôle de Flux/Filtrage Lignes***).
5. Insérer le résultat obtenu dans une base de données que vous créerez au préalable.
6. Insérer le résultat dans une base de données qui va contenir cette fois ci un nouvel attribut Somme = Somme1+Somme2 (**Transformation/Creation d'operations de calcul**)
7. Mettre dans un fichier puis dans une base de données que les clients dont le siège se trouve en Alabama.
8. Nous allons maintenant mettre les clients, et leurs adresses dans deux tables d'une base de données (de votre choix oracle, mysql etc…) avant d'effectuer la jointure !!!