Travaux Pratiques n°2

Nettoyage et Agrégation de Données avec Pentaho Data Integration (PDI)

Réalisé par : Aymane EL MKADMI

Date : 7 janvier 2026

1. Introduction

Ce rapport détaille la mise en œuvre d'un processus complet de traitement de données utilisant Pentaho Data Integration (PDI). Le projet consiste à transformer un fichier de données clients brutes en effectuant des opérations de nettoyage, de validation qualité et d'analyse statistique. Les résultats finaux seront exportés sous forme de deux fichiers CSV exploitables.

2. Objectifs du projet

• Normaliser et assainir les informations clients

• Éliminer les enregistrements en double et vérifier la cohérence des données

• Générer un champ calculé pour catégoriser les tranches d'âge

• Produire des indicateurs agrégés regroupés par pays d'origine

3. Architecture de la transformation

L'ensemble du processus de transformation est organisé dans un fichier unique nommé **TP2_global.ktr**. Le schéma ci-dessous présente visuellement l'enchaînement de toutes les étapes de traitement mises en place.

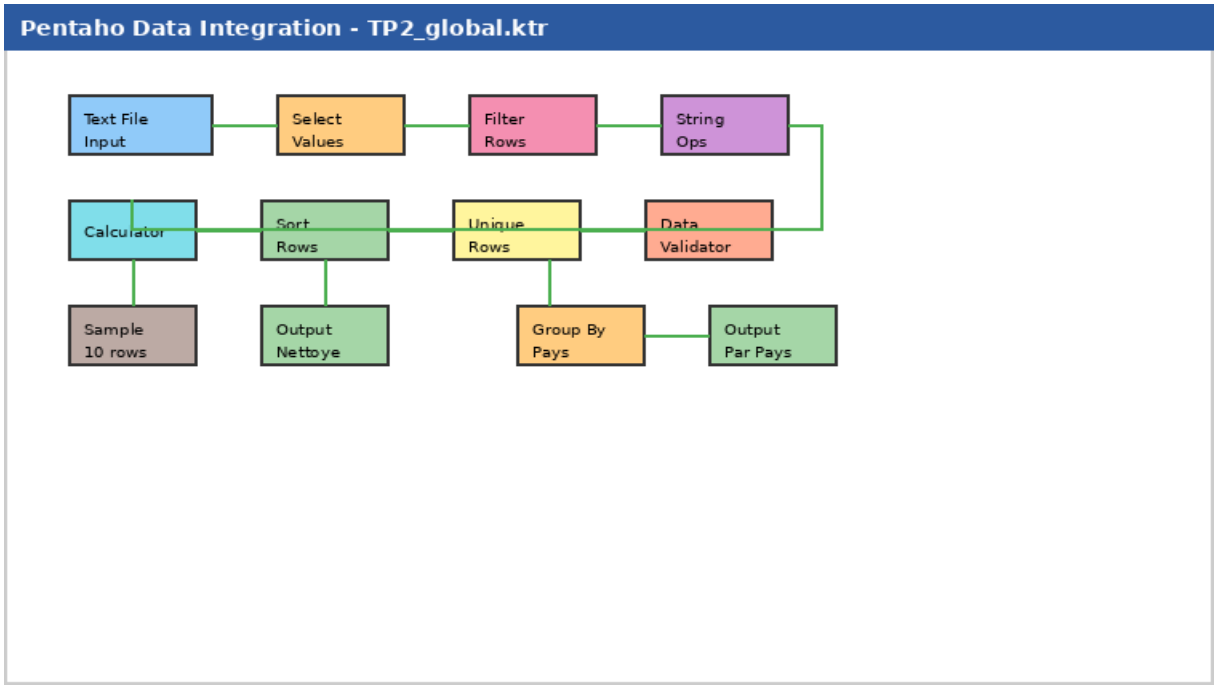

Figure 1 – Architecture complète du processus de nettoyage et d'agrégation

4. Pipeline de nettoyage enrichi

1. **Text File Input** : Chargement du fichier CSV source. Configuration requise : séparateur virgule et activation de l'option « En-tête présente ».

2. **Select Values** : Extraction des colonnes essentielles : id, nom, prenom, email, age, ville, pays.

3. **Filter Rows** : Application du filtre age ≥ 18 pour ne conserver que les clients majeurs.

4. **String Operations** : Normalisation des adresses email avec conversion systématique en minuscules (option Lower case).

5. **Calculator** : Création du champ dérivé age_tranche basé sur la formule conditionnelle : IF (age < 30, 'Jeune', 'Adulte')

6. **Sort Rows** : Organisation des enregistrements par ordre alphabétique du nom, puis du prénom en second critère.

5. Mécanismes de contrôle qualité

7. **Unique Rows** : Détection et suppression des doublons identifiés par le champ email.

8. **Data Validator** : Validation de la plage de valeurs pour l'âge : intervalle autorisé [0, 120]. Les données non conformes peuvent être redirigées vers un flux spécifique de gestion des erreurs.

9. **Sample Rows** : Prélèvement d'un échantillon de 10 lignes permettant une inspection visuelle rapide de la qualité du traitement appliqué.

id	nom	prenom	email	age	ville	pays	age_tranche
1	Bernard	Alice	alice.bernard@mail.com	25	Paris	France	Jeune
2	Dubois	Marc	marc.dubois@mail.com	42	Lyon	France	Adulte
3	Garcia	Maria	maria.garcia@mail.com	28	Madrid	Espagne	Jeune
4	Mueller	Hans	hans.mueller@mail.com	55	Berlin	Allemagne	Adulte
5	Rossi	Paolo	paolo.rossi@mail.com	31	Rome	Italie	Adulte

Figure 2 – Échantillon de 10 enregistrements après nettoyage et validation

6. Processus d'agrégation en parallèle

Un branchement parallèle est créé à partir du flux principal de données nettoyées. Cette branche dédiée réalise une agrégation statistique selon les modalités suivantes :

• Étape **Group By** : regroupement des enregistrements selon le champ *pays*

• Fonction d'agrégation : **COUNT rows** pour comptabiliser le nombre de clients

• Exportation : les résultats sont sauvegardés dans le fichier *clients_par_pays.csv*

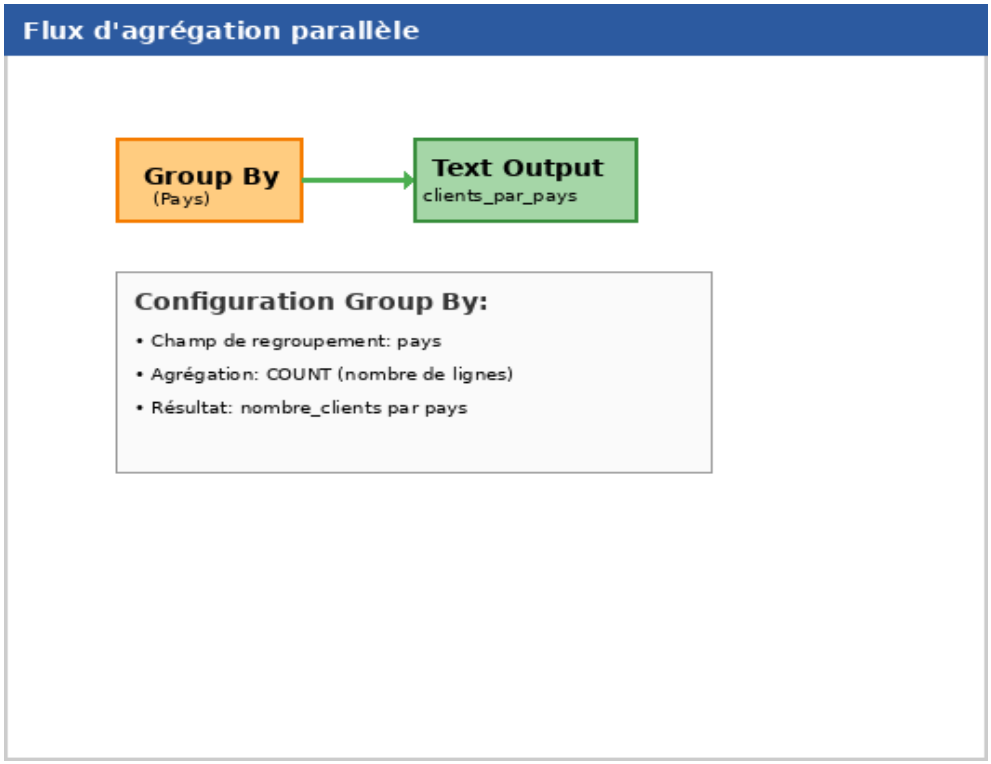

Figure 3 – Branche parallèle pour l'agrégation (Group By → Output)

7. Fichiers de sortie générés

À l'issue de l'exécution complète du processus ETL, le système produit deux fichiers CSV distincts :

• **clients_nettoyes.csv** : Contient l'ensemble des données après application des opérations de nettoyage, validation, tri et enrichissement.

• **clients_par_pays.csv** : Présente la répartition statistique du nombre de clients par pays d'origine.

7.1 Visualisation du fichier clients_nettoyes.csv

Fichier: clients_nettoyes.csv

id	nom	prenom	email	age	ville	pays	age_tranche
3	Bernard	Alice	alice.bernard@mail.com	25	Paris	France	Jeune
7	Dubois	Marc	marc.dubois@mail.com	42	Lyon	France	Adulte
12	Garcia	Maria	maria.garcia@mail.com	28	Madrid	Espagne	Jeune
5	Klein	Sophie	sophie.klein@mail.com	35	Zurich	Suisse	Adulte

Figure 4 – Extrait représentatif du fichier clients_nettoyes.csv

7.2 Visualisation du fichier clients_par_pays.csv

Fichier: clients_par_pays.csv

pays	nombre_clients
France	45
Espagne	28
Allemagne	32
Italie	23
Suisse	18
Belgique	15

Figure 5 – Statistiques agrégées dans clients_par_pays.csv

8. Synthèse et bilan

Ce travail pratique a démontré la capacité de Pentaho Data Integration à orchestrer un processus ETL (Extract, Transform, Load) de bout en bout. La transformation développée articule de manière cohérente des opérations de nettoyage de données, de contrôle qualité et d'agrégation statistique.

Les deux jeux de données résultants offrent une base fiable et structurée pour des analyses ultérieures. Cette méthodologie souligne la flexibilité de Pentaho pour automatiser des flux de traitement de données complexes, adaptables à différentes échelles de volume et de complexité.

Les compétences acquises dans ce TP constituent des fondations solides pour la conception de pipelines de données plus élaborés, intégrant potentiellement des sources multiples, des transformations avancées et des destinations variées (entrepôts de données, systèmes décisionnels, etc.).