

TP2 — Nettoyage et Agrégation de Données avec Pentaho Data Integration (PDI)

Fait par :**Khadija Nachid Idrissi**
Rajae Fdili
Aya Hamim

7 janvier 2026

1 Introduction

Ce rapport présente la réalisation d'un schéma complet de nettoyage et d'agrégation de données avec Pentaho Data Integration (PDI). L'objectif est de traiter un fichier client brut, d'y appliquer des opérations de nettoyage, de contrôle qualité et d'analyse statistique, puis d'exporter les résultats dans deux fichiers CSV exploitables.

2 Objectifs

- Nettoyer et normaliser les données client.
- Supprimer les doublons et valider la cohérence des valeurs.
- Créer une variable dérivée de tranche d'âge.
- Générer des statistiques agrégées par pays.

3 Schéma global de la transformation

Le flux de transformation complet est illustré dans la figure suivante. Il regroupe l'ensemble des étapes du TP2 dans un seul fichier `TP2_global.ktr`.

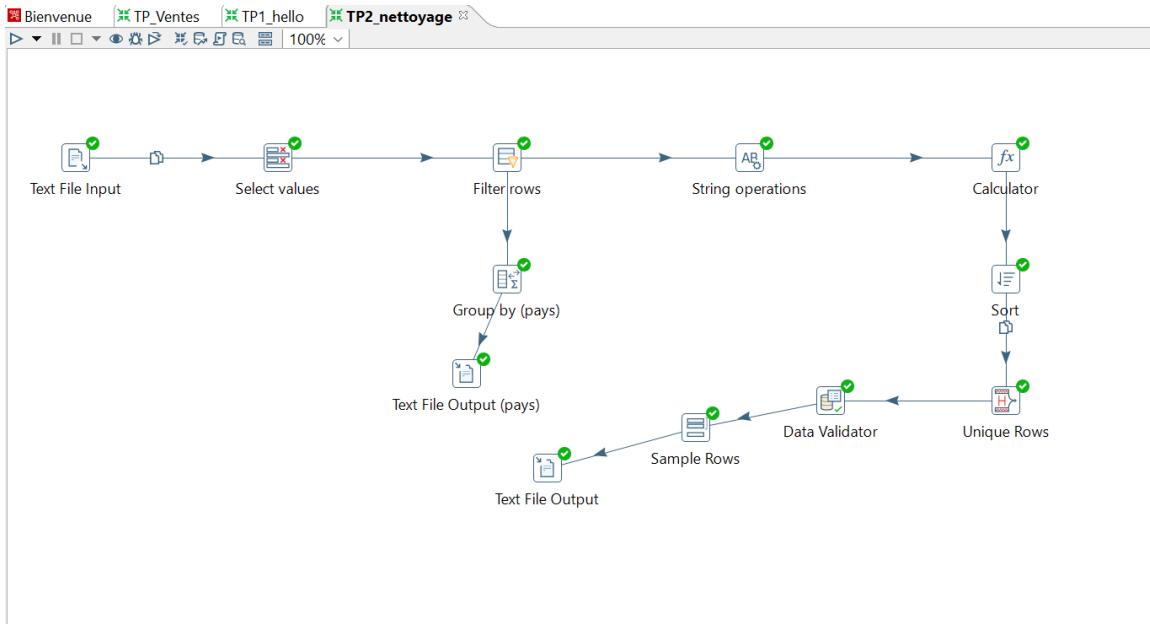


FIGURE 1 – Schéma complet du nettoyage et de l’agrégation sous Pentaho

4 Étapes du nettoyage enrichi

1. **Text File Input** : import du fichier CSV contenant les informations clients. Le séparateur est la virgule, et l’option « En-tête présente » est activée.
2. **Select Values** : sélection des colonnes pertinentes (`id`, `nom`, `prenom`, `email`, `age`, `ville`, `pays`).
3. **Filter Rows** : condition `age > 18` pour exclure les mineurs.
4. **String Operations** : conversion des adresses email en minuscules (option `Lower case`).
5. **Calculator** : ajout du champ `age_tranche` via la formule :

```
IF (age < 30, 'Jeune', 'Adulte')
```

6. **Sort Rows** : tri des clients par `nom` puis `prenom`.

5 Contrôles qualité

7. **Unique Rows** : suppression des doublons basés sur `email`.
8. **Data Validator** : vérifie que `age >= 0` et `age <= 120`. Les enregistrements invalides peuvent être redirigés vers un flux d’erreur si nécessaire.
9. **Sample Rows** : échantillonnage de 10 lignes pour un contrôle visuel des résultats.

Lignes de l'étape: Text File Input (10 lignes)

#	id	nom	prenom	email	age	ville	pays
1	1	Dupont	Jean	jean.dupont@example.com	34	Paris	Bresil
2	2	Martin	Claire	claire.martin@example.com	28	Lyon	Turquie
3	3	Marques	Michele	michele.marques@example.com	29	Potierne	Russie
4	4	Moulin	Helene	helene.moulin@example.com	42	Sainte AlexBourg	Coree du Sud
5	5	Bourdon	Marcelle	marcelle.bourdon@example.com	29	Paul	Maroc
6	6	Letellier	Chantal	chantal.letellier@example.com	50	Sainte Catherine	Afrique du Sud
7	7	Da Silva	Amelie	amelie.da.silva@example.com	68	Delaunay	Russie
8	8	Denis	Adelaide	adelaide.denis@example.com	43	Garnierne	Turquie
9	9	Blanc	Anne	anne.blanc@example.com	48	Bertin	Danemark
1.	10	Renard	Charles	charles.renard@example.com	28	Louisnec	Russie

FIGURE 2 – Aperçu de l'échantillon de 10 lignes nettoyées et validées

6 Agrégation parallèle

10. Group By : un flux parallèle est ajouté à partir du flux nettoyé. Le regroupement s'effectue par le champ **pays**, avec la fonction d'agrégation **COUNT rows**. Le résultat est exporté dans un second fichier CSV contenant le nombre de clients par pays.

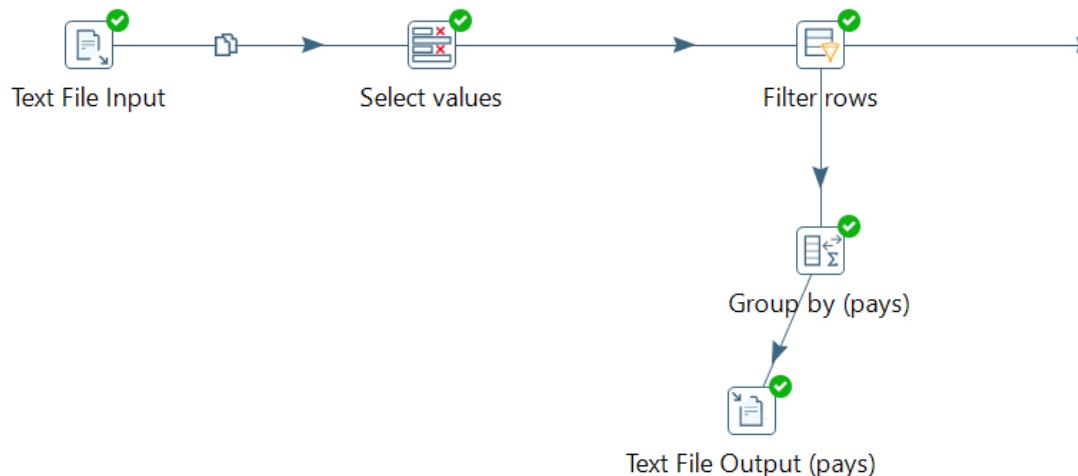


FIGURE 3 – Flux parallèle d'agrégation (Group by → Output)

7 Résultats finaux

Après exécution de la transformation complète, deux fichiers CSV sont générés :

- **clients_nettoyes.csv** : contient les données nettoyées, triées, validées et enrichies.
- **clients_par_pays.csv** : contient le nombre de clients regroupés par pays.

7.1 Aperçu du fichier clients_nettoyes.csv

#	id	nom	prenom	email	age	ville	pays	age_tranche
1	2577	Adam	Alice	alice.adam@example.com	52	Dumont-les-Bains	Afrique du Sud	Adulte
2	4401	Adam	Bernadette	bernadette.adam@example.com	34	Sainte Claudine	Espagne	Adulte
3	1041	Adam	Christine	christine.adam@example.com	39	Gaudin	Grece	Adulte
4	2783	Adam	Gabriel	gabriel.adam@example.com	46	Grenier	Grece	Adulte
5	2393	Adam	Valentine	valentine.adam@example.com	25	Le Gall	Russie	Jeune
6	3731	Adam	Xavier	xavier.adarn@example.com	66	Marchand	Espagne	Adulte
7	3815	Adam	Zacharie	zacharie.adam@example.com	39	Guyon	Etats-Unis	Adulte
8	3369	Albert	Alex	alex.albert@example.com	34	Laroche-les-Bains	Italie	Adulte
9	877	Albert	Antoine	antoine.albert@example.com	60	Larocheboeuf	Japon	Adulte
1.	41	Albert	Emilie	emilie.albert@example.com	24	Grenier-sur-Bousquet	Bresil	Jeune
1.	3165	Albert	Gregoire	gregoire.albert@example.com	42	Leclercnec	Mexique	Adulte
1.	2157	Albert	Hugues	hugues.albert@example.com	65	Sainte Vincentnec	Suede	Adulte
1.	2323	Albert	Marcel	marcel.albert@example.com	53	Davidboeuf	Australie	Adulte
1.	279	Albert	Pauline	pauline.albert@example.com	41	Baudry	Suede	Adulte
1.	2351	Albert	Thibaut	thibaut.albert@example.com	24	Antoine	Turquie	Jeune
1.	1081	Albert	Valentine	valentine.albert@example.com	33	Begue	Norvege	Adulte
1.	2383	Alexandre	Anais	anais.alexandre@example.com	61	Francoisdan	Canada	Adulte
1.	4783	Alexandre	Corinne	corinne.alexandre@example.com	58	Blanchard-sur-Thierry	Inde	Adulte
1.	1287	Alexandre	Maurice	maurice.alexandre@example.com	36	Roussel	Mexique	Adulte
2.	2865	Allain	Frederic	frederic.allain@example.com	68	Robinboeuf	Tunisie	Adulte
2.	423	Allain	Ines	ines.allain@example.com	31	Gregoire	Portugal	Adulte
2.	187	Allain	Jules	jules.allain@example.com	41	Nicolas-sur-Lacroix	Mexique	Adulte

FIGURE 4 – Extrait du fichier **clients_nettoyes.csv**

7.2 Aperçu du fichier clients_par_pays.csv

#	pays	nombre_clients
1	Etats-Unis	93
2	Allemagne	72
3	Bresil	67
4	Turquie	90
5	Mexique	78
6	Danemark	80
7	Canada	79
8	Italie	102
9	Argentine	82
1.	Australie	80
1.	Chine	88
1.	Suisse	86
1.	Maroc	84
1.	France	86
1.	Tunisie	88
1.	Algerie	84
1.	Pologne	92
1.	Egypte	74
1.	Japon	69
2.	Pays-Bas	77
2.	Suede	90
2.	Afrique du Sud	94
2.	Espagne	80

FIGURE 5 – Extrait du fichier **clients_par_pays.csv**

8 Conclusion

Ce TP a permis de mettre en place un processus complet d'ETL (Extraction, Transformation et Chargement) sous Pentaho. La transformation combine des opérations de nettoyage, de validation et d'agrégation pour produire deux jeux de données fiables et exploitables. Cette approche illustre la puissance et la souplesse de Pentaho pour automatiser des traitements de données à grande échelle.