

Travaux Pratiques n°1

Prise en main de Pentaho Data Integration

Réalisé par : Aymane EL MKADMI

Date : 7 janvier 2026

1. Objectifs du laboratoire

Ce laboratoire vise à acquérir les compétences fondamentales dans l'utilisation de Pentaho Data Integration (PDI). Vous allez développer votre première transformation ETL qui permettra d'importer des données depuis un fichier CSV, d'effectuer une sélection ciblée de champs, puis de sauvegarder le résultat obtenu dans un nouveau document.

2. Données sources

Le dataset initial *clients.csv* contient l'ensemble des informations suivantes :

id	nom	prenom	email	age	ville	pays
1	Dupont	Jean	jean.dupont@example.com	34	Paris	France
2	Martin	Claire	claire.martin@example.com	28	Lyon	France

Notre mission consiste à extraire exclusivement quatre colonnes : **id**, **nom**, **prenom** et **email**.

3. Étapes de réalisation

3.1 Initialisation du projet

1. Lancez Pentaho Data Integration
2. Accédez au menu *Fichier* puis sélectionnez *Nouvelle → Transformation*
3. Sauvegardez immédiatement votre travail sous le nom **TP1_hello.ktr**

3.2 Assemblage des composants

La transformation nécessite trois éléments essentiels qu'il faut configurer et connecter :

- **Text file input** : Permet la lecture du fichier source. Assurez-vous d'activer l'option *Header row present* et de définir la virgule comme séparateur de champs.
- **Select values** : Situé dans la catégorie *Altération structure flux*, ce composant permet de filtrer les colonnes à conserver (id, nom, prenom, email uniquement).
- **Text file output** : Génère le fichier résultat *clients_hello.csv* contenant les données transformées.

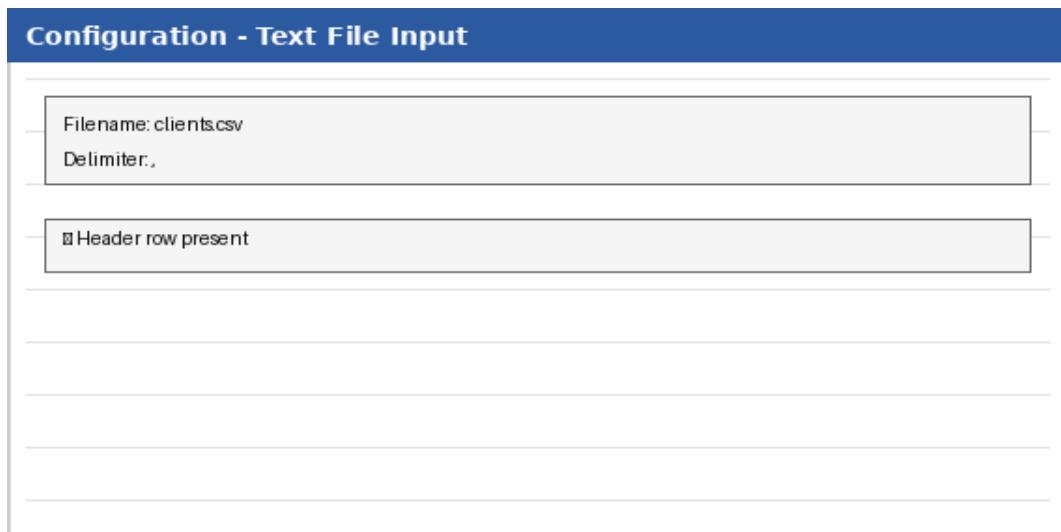


Figure 1 – Configuration du composant Text File Input

Reliez ces trois composants dans l'ordre logique en utilisant des *hops* (connexions représentées par des flèches vertes).

3.3 Lancement de la transformation

Une fois tous les éléments configurés et connectés, cliquez sur le bouton d'exécution pour lancer le processus de transformation.

4. Paramètres critiques

Certaines configurations sont essentielles pour garantir le bon fonctionnement de votre transformation :

1. Dans **Text file input** : activez impérativement *Header row present*, choisissez l'encodage UTF-8, définissez la virgule comme délimiteur et les guillemets doubles si vos données en contiennent.
2. Utilisez la fonction **Get fields** pour que PDI détecte automatiquement la structure de vos données (noms de colonnes et types de données).
3. Vérifiez visuellement que toutes les connexions (flèches vertes) sont bien établies entre les composants. L'absence de connexion empêchera totalement le flux de données.

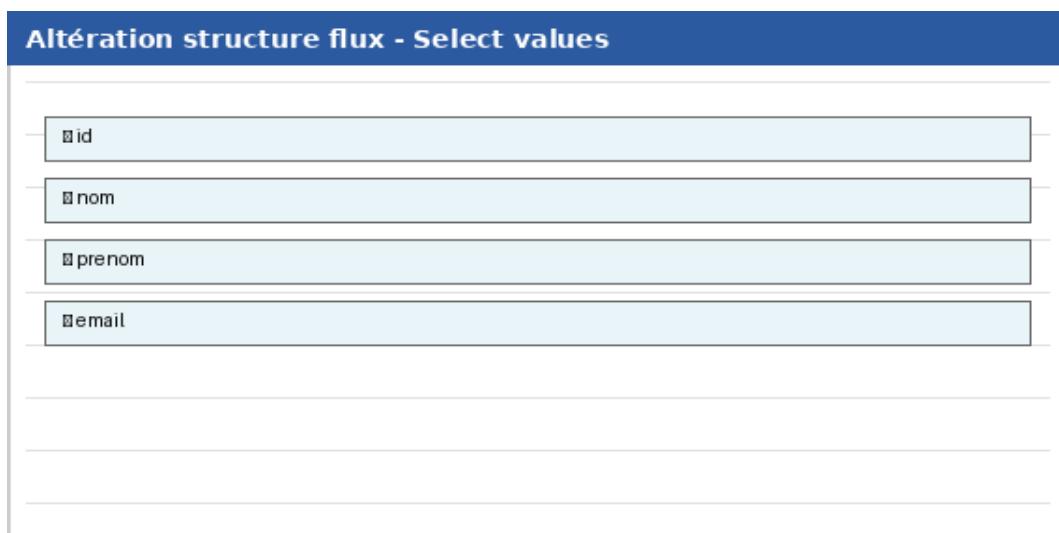


Figure 2 – Configuration du composant Altération structure flux pour sélectionner les colonnes

5. Validation des résultats

Après l'exécution réussie de la transformation, le fichier de sortie *clients_hello.csv* doit contenir exclusivement les quatre colonnes sélectionnées :

id	nom	prenom	email
1	Dupont	Jean	jean.dupont@example.com
2	Martin	Claire	claire.martin@example.com

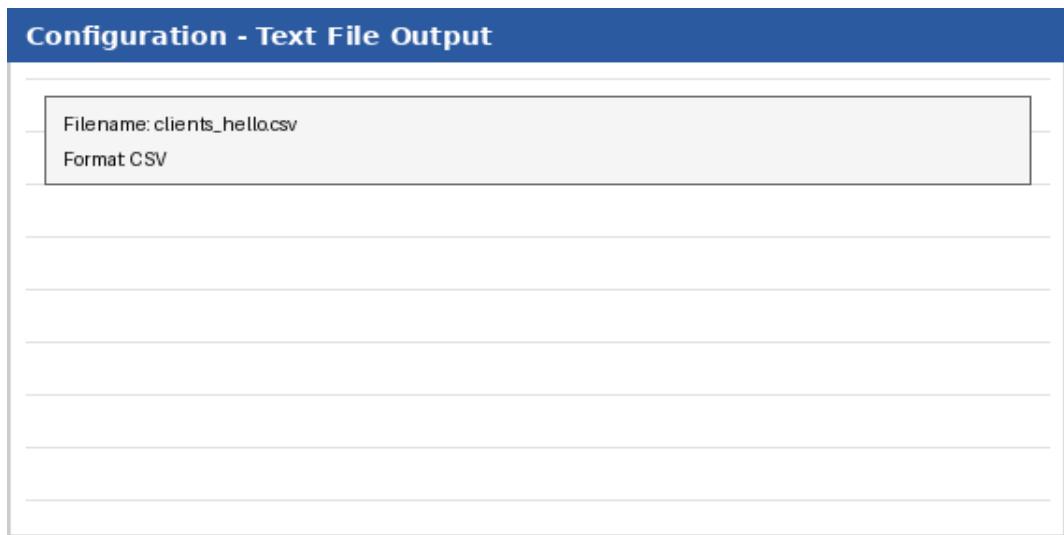


Figure 3 – Configuration du composant Text File Output

Le processus s'est déroulé conformément aux attentes : les données ont été correctement filtrées et seuls les champs pertinents ont été conservés dans le document final.

Flux complet de transformation

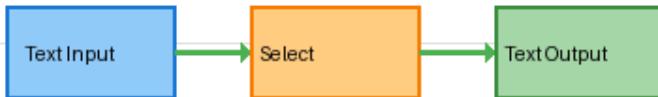


Figure 4 – Flux complet de la transformation dans Pentaho PDI

6. Synthèse et perspectives

Ce premier exercice pratique vous a permis de vous approprier les fonctionnalités de base de Pentaho Data Integration. Vous maîtrisez maintenant les opérations fondamentales : import de fichiers CSV, sélection ciblée de colonnes et export des données transformées.

Cette transformation simple constitue le socle sur lequel vous pourrez construire des processus ETL (Extract, Transform, Load) bien plus élaborés, intégrant des jointures, des agrégations, des enrichissements de données et des chargements vers diverses destinations (bases de données, API, etc.).