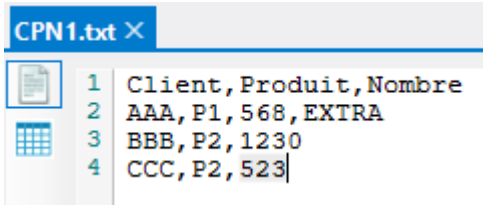
Utiliser Integrator pour nettoyer manuellement les données créées

Description

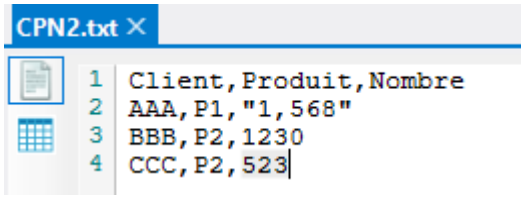
Supposons que nous ayons plusieurs fichiers texte en entrée dans un script Integrator et n'ayant pas strictement la même structure. Les données peuvent être nettoyées en modifiant les propriétés d'entrée et en ajoutant des calculs au sein d'Integrator.

Le premier fichier texte a le contenu suivant :



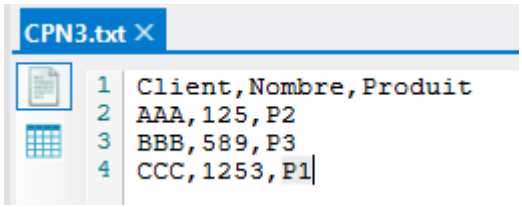
On constate que sur la première ligne de donnée, un champ supplémentaire est présent.

Le deuxième fichier texte possède le contenu suivant :



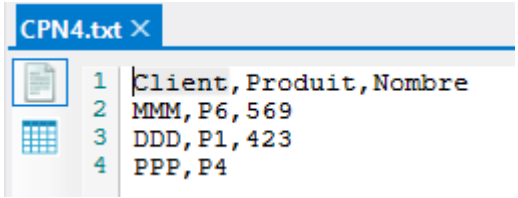
Sur la première ligne de donnée, le séparateur de milliers pour les nombres est la virgule.

Le troisième fichier texte contient les données suivantes :



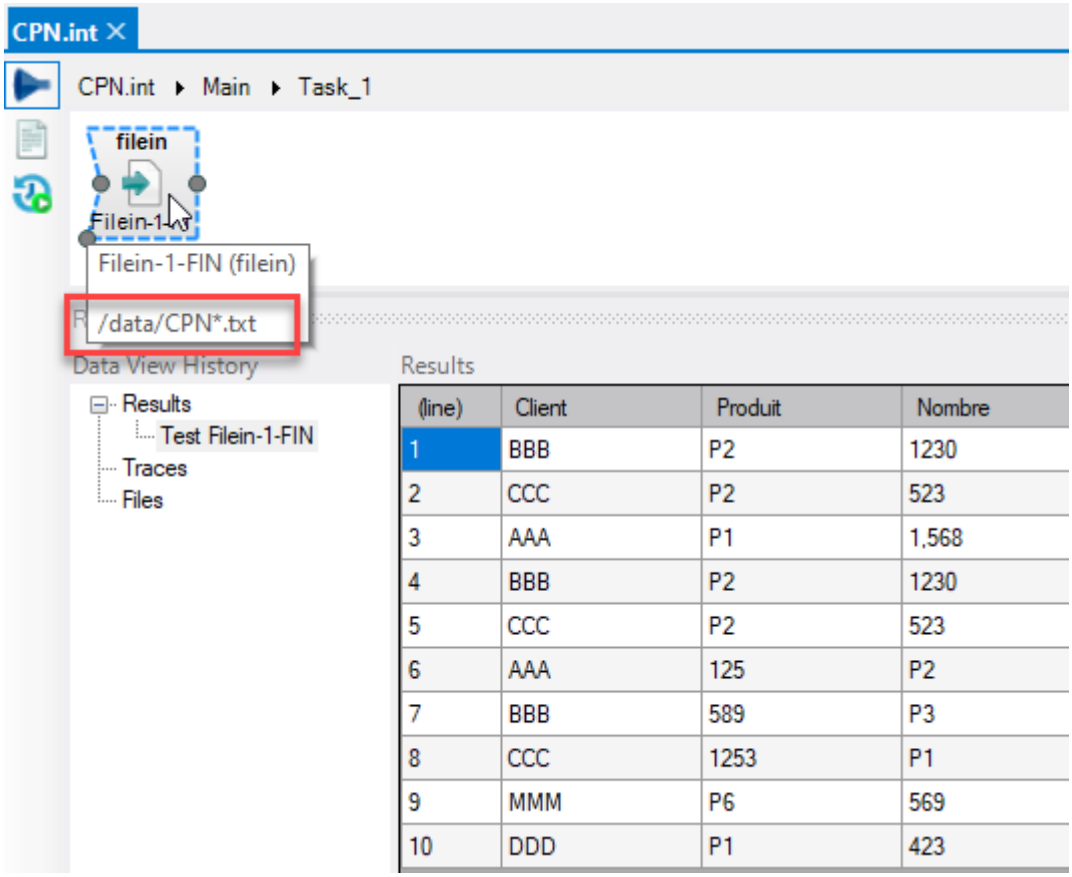
L'ordre d'apparition des colonnes est différent des deux autres fichiers.

Enfin un quatrième fichier est peuplé de la manière suivante :



On constate qu'il manque une valeur de champ sur la dernière ligne.

L'utilisation d'un script Integrator permettant la consolidation des données des 4 fichiers précédents donne ceci :



Results

(line)	Client	Produit	Nombre
1	BBB	P2	1230
2	CCC	P2	523
3	AAA	P1	1,568
4	BBB	P2	1230
5	CCC	P2	523
6	AAA	125	P2
7	BBB	589	P3
8	CCC	1253	P1
9	MMM	P6	569
10	DDD	P1	423

On constate que seules ***10*** lignes sont affichées alors que la somme des lignes des 4 fichiers donne ***12***.

Certaines lignes sont donc ignorées.

Pour rétablir les lignes manquantes, la première modification à effectuer est de paramétrer l'option ***Ignore_Extra_Column*** à ***true*** au niveau de l'objet d'entrée Filein.

La deuxième modification à effectuer est de paramétrer l'option ***Ignore_Line_End*** à ***true*** au niveau de l'objet d'entrée Filein.

Une fois les options précitées modifiées, un test sur l'objet Filein donne ceci :

Results

(line)	Client	Produit	Nombre
1	AAA	P1	568
2	BBB	P2	1230
3	CCC	P2	523
4	AAA	P1	1,568
5	BBB	P2	1230
6	CCC	P2	523
7	AAA	125	P2
8	BBB	589	P3
9	CCC	1253	P1
10	MMM	P6	569
11	DDD	P1	423
12	PPP	P4	

On a bien maintenant 12 lignes en sortie.

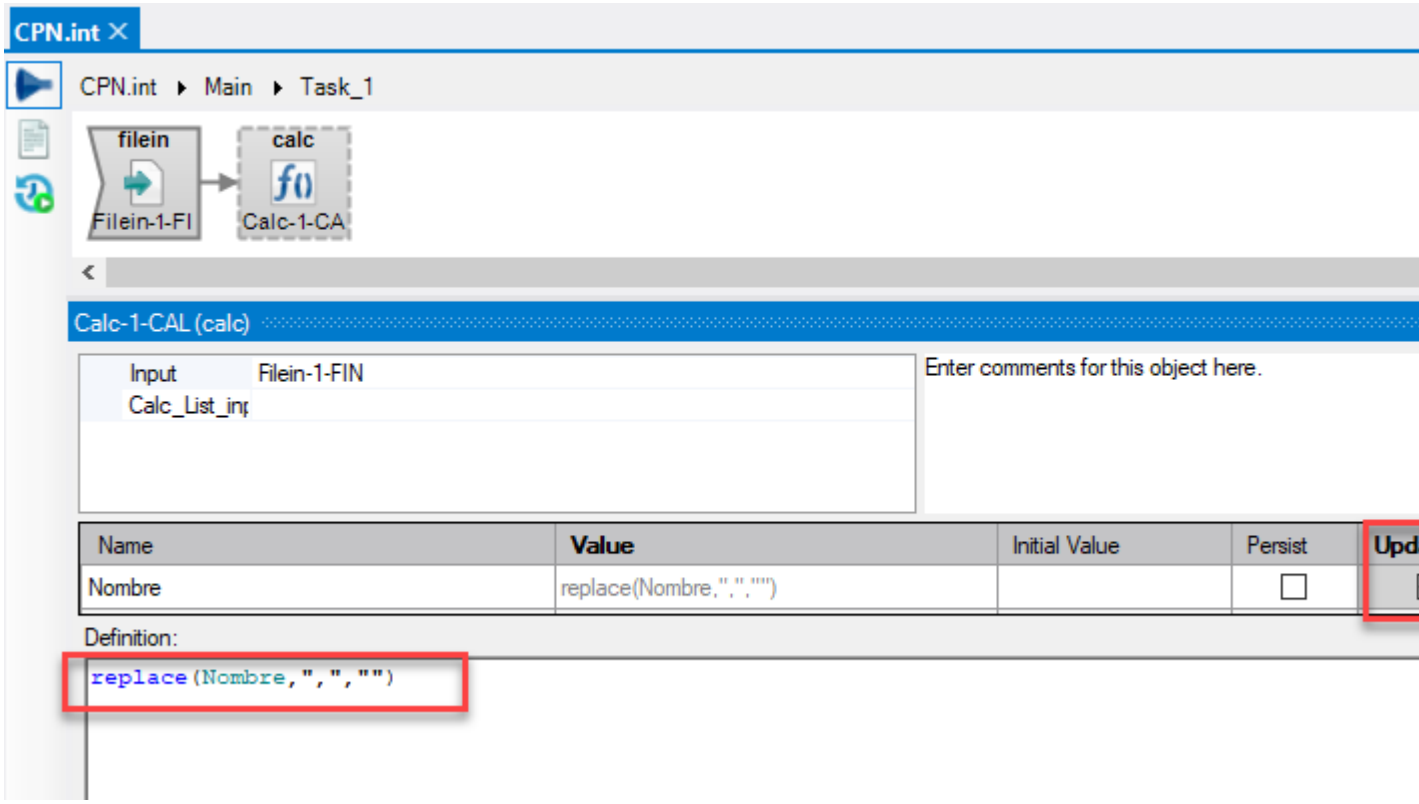
On a cependant un problème au niveau des valeurs des champs. Concernant les lignes 7 à 9 on a une permutation des valeurs des champs ***Produit*** et ***Nombre***.

La modification de l'option ***Union*** à ***true*** au niveau de l'objet d'entrée Filein, permet d'obtenir une alimentation correcte des deux champs :

(line)	Client	Produit	Nombre
1	AAA	P1	568
2	BBB	P2	1230
3	CCC	P2	523
4	AAA	P1	1,568
5	BBB	P2	1230
6	CCC	P2	523
7	AAA	P2	125
8	BBB	P3	589
9	CCC	P1	1253
10	MMM	P6	569
11	DDD	P1	423
12	PPP	P4	

Il reste à modifier le formatage du nombre sur la ligne 4 en supprimant le séparateur de milliers pour passer de la valeur ***1.568*** à ***1568***.

L'ajout d'un objet Calc au flux de données avec la formule de calcul suivante :



et l'option ***Update*** cochée donne ceci :

Results

(line)	Client	Produit	Nombre
1	AAA	P1	568
2	BBB	P2	1230
3	CCC	P2	523
4	AAA	P1	1568
5	BBB	P2	1230
6	CCC	P2	523
7	AAA	P2	125
8	BBB	P3	589
9	CCC	P1	1253
10	MMM	P6	569
11	DDD	P1	423
12	PPP	P4	

Tags

1. Data Integrator
2. script
3. Visual Integrator