Pentaho Data Integration

version 2.5.0

Foire aux questions

traduction française *en cours*
Partie 1/6

en cas de doute se référer au document original en anglais

Table des matières

1 Preface..... 4

2 Les questions des utilisateurs débutants..... 5

2.1 Quelle est la différence entre une transformation (ndt en anglais transformation) et une tâche (ndt en anglais job) ?..... 5

2.2 Règle sur le mélange des type de lignes sur un lien dans une transformation..... 5

2.3 Nom de champ en double au sein d'une transformation..... 5

2.4 Au sujet des chaînes de caractères vides ou nulles (NULL)..... 6

2.5 Comment copier/dupliquer un champ au sein d'une transformation ?..... 6

2.6 Comment réaliser une jointure au sein d'une base de données avec PDI ?..... 7

2.7 Comment sérialiser les transformations ?..... 7

3 Further User questions..... 8

3.1 Strings bigger than defined String length..... 8

3.2 Decimal point doesn't show in .csv output..... 8

3.3 Function call returning boolean fails in Oracle..... 9

3.4 Difference between variables/arguments in launcher..... 9

3.5 How to use database connections from repository..... 9

3.6 On inserting booleans into a MySql database..... 10

3.7 Calculator ignores result type on division..... 10

3.8 HTTP Client Step questions..... 10

3.8.1 The HTTP client step doesn't do anything..... 10

3.8.2 The HTTP client step and SOAP..... 11

3.9 Javascript questions..... 12

3.9.1 How to check for the existence of fields in rows..... 12

3.9.2 How to add a new field in a row..... 12

3.9.3 How to replace a field in a row..... 12

3.9.4 How to create a new row..... 13

3.9.5 How to use something as nvl in javascript?..... 13

3.9.6 Example of how to split fields..... 13

3.10 Shell job entry questions..... 14

3.10.1 How to check for the return code of a shell script/batch file..... 14

4 Twilight user-development questions..... 15

4.1 Implement connection type as a variable parameter 15

4.2 Implement “on the fly DDL” creation of tables, ... 15

4.3 Implement a step that shows a dialog and asks parameters..... 15

4.4 Implement serialization of transformations using reflection..... 16

4.5 Implement retry on connection issues..... 16

4.6 Implement GUI components in a transformation or job..... 17

5 Development questions..... 18

5.1 Priority on development..... 18

5.1.1 Correctness/Consistency..... 18

5.1.2 Backwards compatibility..... 18

5.1.3 Speed..... 18

5.1.4 User friendliness..... 18

5.1.4.1 Division of functionality in steps and job entries..... 18

5.1.4.2 Rows on a single hop have to be of the same structure..... 18

5.1.4.3 Null and “” are the same in PDI 18

5.1.4.4 On converting data to fit the corresponding Metadata..... 19
5.2 On logging in steps in Pentaho Data Integration..... 20
5.3 On using XML in Pentaho Data Integration..... 20
5.4 On using I18N in PDI..... 21
5.5 On reformatting source code..... 21
5.6 On using Subversion..... 22
6 Success factors of PDI..... 23
6.1 Modular design..... 23

1 PRÉFACE

Ce document contient la “Foire aux questions” fréquemment posées au sujet de “Pentaho Data Integration”, anciennement connu sous le nom de Kettle. Les questions et réponses de ce document sont principalement issues des questions posées sur le forum de discussion. Si une question se trouve ici, il est probablement inutile de la poser de nouveau sur le forum de discussion. Cependant, si une réponse de ce document vous apparaît ,ne pas être claire, merci de faire référence à cette foire aux questions dans le forum de discussion.

2 LES QUESTIONS DES UTILISATEURS DÉBUTANTS

2.1 Quelle est la différence entre une transformation (ndt en anglais transformation) et une tâche (ndt en anglais job) ?

Q: Dans Spoon, je peux fabriquer des tâches et des transformations. Quelle est la différence entre les deux ?

A: Les transformations servent à déplacer et transformer des lignes depuis un flux source de données vers un flux cible. On utilise les tâches à un niveau conceptuel plus élevé dans le contrôle des processus, par exemple pour exécuter des transformations, envoyer un courriel sur erreur, charger un fichier etc.

2.2 Règle sur le mélange des type de lignes sur un lien dans une transformation

Q: Dans les documentations, je lis que les types de lignes ne peuvent pas être mélangés; qu'est-ce que cela signifie ?

A: Ne pas mélanger les lignes signifie que chaque ligne qui est envoyée sur un même lien doit avoir la même structure : les mêmes noms de champs, types de champs, ordre des champs. Si vous souhaitez faire quelque chose du genre « ajouter un champ optionnel si tel condition est vraie ou fausse, alors », cela ne fonctionnera pas (parce que vous récupérerez des type de lignes différents selon la condition). Vous pouvez basculer en « Mode sécurisé » pour vérifier explicitement cette l'identité des types au démarrage.

La vérification des types de lignes est automatiquement réalisée depuis la version 2.5.0 lors de l'étape de design/vérification. Cependant, le « mode sécurisé » doit être activé pour que cette vérification se fasse dès le démarrage.

De la même façon qu'une ligne d'une table de base de données aura toujours la même structure, vous ne pourrez pas enregistrer plusieurs types de lignes dans un flux PDI.

Théoriquement, la raison provient de ce que PDI veux être capable de réaliser des transformations uniformes et consistantes sur vos données. Se doter de type de lignes variables rendrait beaucoup plus complexe ces transformations.

Techniquement, la plupart des étapes d'une transformation sont optimisées (en utilisant des techniques d'identifiant et d'indexation). Se doter de types de lignes variables rendrait cette optimisation impossible.

2.3 Nom de champ en double au sein d'une transformation

Q: Puis-je avoir deux mêmes noms de champ au sein d'une même transformation ?

A: Vous ne pouvez pas. PDI vous avertira de cette impossibilité dans la plupart des cas de doublons sur le nom de champs. Avant la version 2.5.0, vous pouviez forcer cette contrainte, mais seulement la première valeur des champs en doublon était utilisée.

2.4 Au sujet des chaînes de caractères vides ou nulles (NULL)

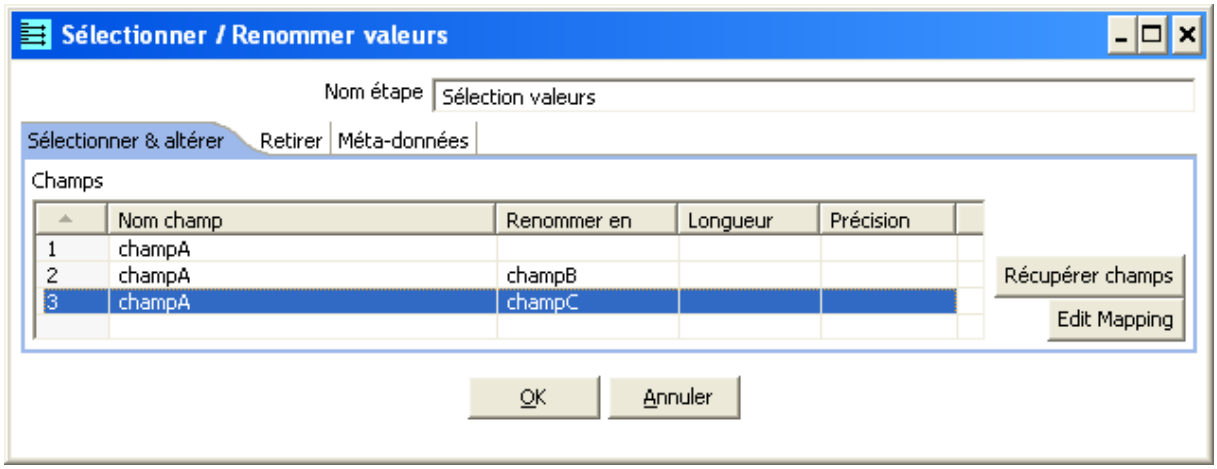
PDI suit Oracle dans son utilisation des chaînes de caractères vides ou nulles : il considère que c'est la même chose. Si vous trouvez une étape qui ne suit pas cette convention, signalez-le. C'est probablement un bogue.

2.5 Comment copier/dupliquer un champ au sein d'une transformation ?

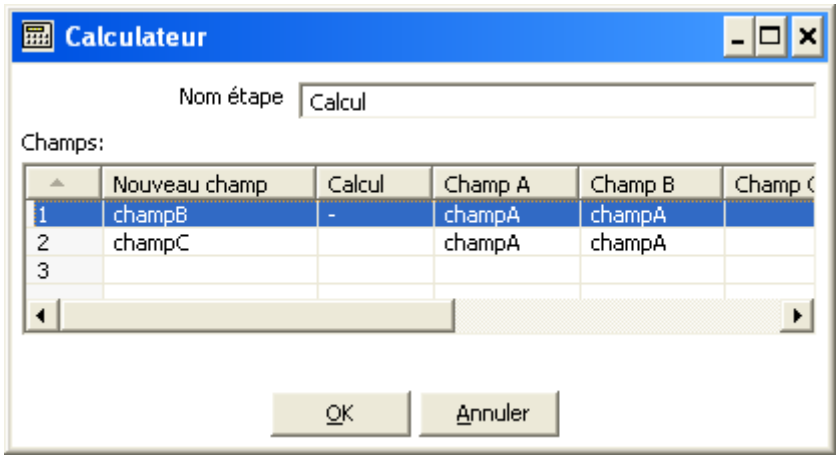
Q: Comment copier/dupliquer un champ au sein d'une transformation ?

A: Plusieurs solutions existent :

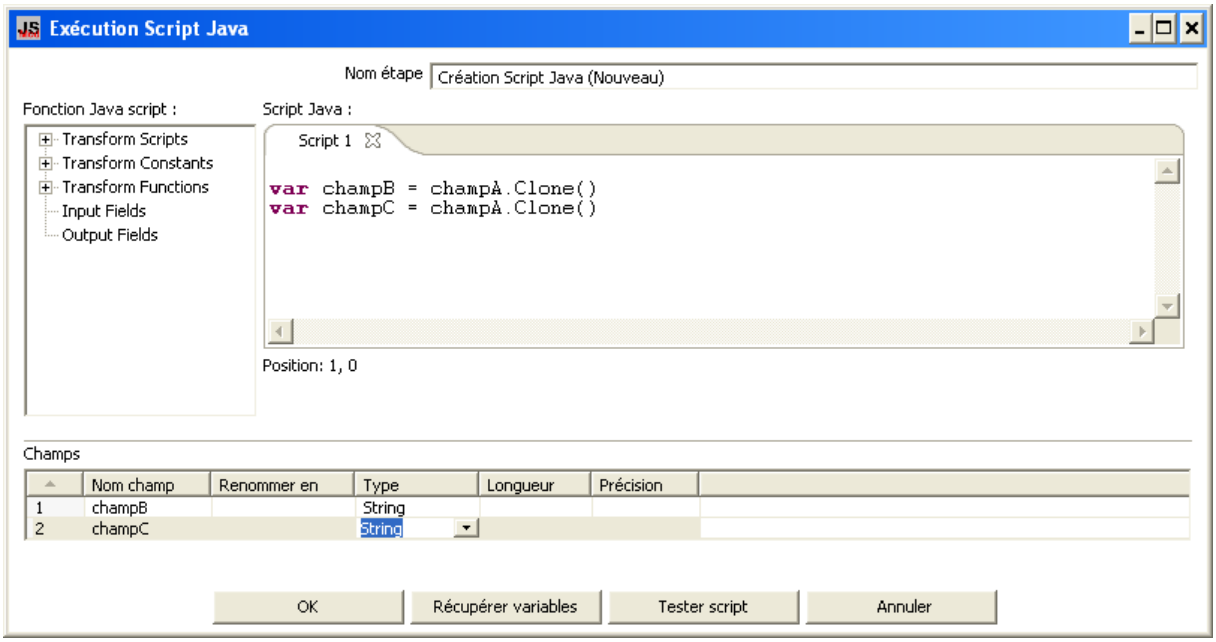
1) Utiliser une étape « Sélectionner valeurs » en renommant un champ sélectionné. Le champ original sera alors dupliqué avec un nouveau nom. On aura par exemple :
Cela dupliquera le champA en champB et champC



2) Utiliser une étape « Calculatrice » et se servir de l'opération NLV(A,B) comme calcul (non disponible en version 2.4.5) comme suit :
Cela aura le même effet que la première solution. Cela dupliquera le champA en champB et champC



3) Utiliser une étape « JavaScript » pour copier le champ
Cela aura le même effet que les précédentes solutions : 3 champs dans le flux de sortie qui sont des copies les uns des autres :champA, champB, champC.



2.6 Comment réaliser une jointure au sein d'une base de données avec PDI ?

Q: Comment réaliser une jointure au sein d'une base de données avec PDI ?

A: Si vous désirez réaliser une jointure de deux tables provenant d'une même base de données, alors, vous utiliserez l'étape « Entrée table » et exécuterez une instruction SQL de jointure. Avec une seule base de données, c'est la méthode la plus rapide.

Si vous désirez réaliser une jointure de deux tables qui ne proviennent pas de la même base de données, vous pouvez utiliser une étape « Jointure Base de données » mais vous devez vous rendre compte que PDI exécutera une instruction pour chacune des ligne en entrée de la jointure. Vous pourrez trouver des indications supplémentaires dans les « trucs et astuces hebdomadaires » sur le site de Pentaho (http://kettle.pentaho.org/tips/).

2.7 Comment sérialiser les transformations ?

Q: Toutes les étapes, par défaut, tournent en parallèle au sein d'une transformation. Comment puis-je faire pour qu'une ligne soit complètement traitée avant que PDI commence à traiter la ligne suivante ?

A: Ce n'est pas possible, PDI est construit sur cette notion de traitement en parallèle. Il faudrait changer d'architecture pour permettre une « sérialisation » des traitements. Cela induirait un temps de traitement très lent.

3 QUESTIONS D'AUTRES UTILISATEURS

3.1 Chaîne de caractères plus longue que la longueur définie

Q: J'essaie de limiter une chaîne de caractères avec le paramètre « longueur » mais des chaînes plus longues passent au travers de mon étape et cela cause une erreur SQL par la suite. Le programme n'est-il pas censé tronquer la chaîne de caractères à la longueur saisie ? Ou dois-je tronquer moi-même la chaîne de caractères avant insertion ? Si oui, quel est la meilleure façon de faire cela ?

A: Une transformation fonctionne sur les méta-données. Au début de la conception de Kettle, le concepteur a pris le parti de travailler sur les méta-données et non les données elles-mêmes. Ainsi, l'utilisateur n'est pas toujours arrété dans son cheminement par des messages d'erreur sur le typage des données. Cependant, si vous désirez tronquer un champ, vous pourrez le faire en écrivant un script javascript comme celui ci-dessous :

```
var string = originalfield.getString();
if ( string.length() > 50 )
{
    string = string.substring(0,50);
}
```

3.2 Le point décimal n'est pas affiché dans la sortie CSV (ndt fichier délimité par des virgules)

Q: Dans une table de ma base de données, j'ai une colonne contenant un décimal :

```
100.23
100.20
100.00
100.00
```

En me servant d'un étape « sortie vers fichier » et en utilisant l'extension CSV, lorsque j'ouvre ce fichier dans mon tableur (MS Excel), j'obtiens ceci :

```
100.23
100.20
100
100
```

Q: Comment obtenir systématiquement les décimales (“.00”) même si elles sont nulles ?

A: En premier lieu, il vous faut utiliser un format "#.00" pour ce champ dans le paramétrage de votre étape. Vous obtiendrez une précision de deux décimales, quelles qu'elles soient, ce que vous pourrez vérifier en ouvrant votre fichier dans un programme d'édition de fichier (Bloc Notes par exemple).

Si vous ouvrez le fichier d'extension csv avec MS Excel, les deux décimales lorsqu'elles sont nulle n'apparaîtront peut-être pas car MS Excel utilise par défaut un formatage qui cache les décimales nulles. Utilisez donc l'étape « Sortie vers excel » qui, en version 2.4.0, permet des formatages simples pour Excel.