

1 Objectif

Description succincte de Pentaho Data Integration Community Edition (Kettle).

L'informatique décisionnelle (« [Business Intelligence](#) – BI » en anglais, ça fait tout de suite plus *glamour*) fait référence à « l'exploitation des données de l'entreprise dans le but de faciliter la prise de décision ». Des suites logicielles se proposent de prendre en charge le processus complet. J'ai choisi de mettre en avant la suite Open Source [Pentaho](#), mais les principes énoncés sont valables pour la grande majorité des logiciels du domaine.

La suite Pentaho est composée d'une série d'outils associés à chaque étape de la BI. L'outil d'[intégration des données](#) se charge de puiser les informations dans les différentes sources de l'entreprise, de les fusionner (intégrer), de les nettoyer. L'objectif est de nourrir l'entrepôt de données de l'entreprise. On parle d'outil ETL (Extraction, Transformation and Loading). [Pentaho Analysis](#) est un outil OLAP (Online Analytical Processing). Son rôle est d'explorer les données en utilisant des croisements à plusieurs niveaux afin de mettre en avant les informations les plus pertinentes (je simplifie à l'extrême là). [Pentaho Dashboards](#) et [Pentaho Reporting](#) sont des outils qui permettent de produire, respectivement, des tableaux de bords et des rapports destinés à rendre compte de l'activité de l'entreprise. Ils ont tout deux singulièrement contribué au succès de l'informatique décisionnelle en montrant aux décideurs qu'il était possible, à partir d'informations de toute manière déjà disponibles dans le systèmes d'information existants, de produire des indicateurs simples et pertinents pour suivre les performances de l'entreprise dans le temps. [Pentaho Data Mining](#) (basé sur Weka) enfin permet d'approfondir l'analyse exploratoire en s'appuyant sur les techniques de fouille de données.

Il existe deux versions de Pentaho. L'édition entreprise est payante, elle donne accès à une assistance. Je ne l'ai pas testée. La « [Community Edition](#) » (Pentaho CE) est téléchargeable librement. Elle est développée et maintenue par une communauté de développeurs. Je ne situe pas bien les différences entre les deux versions. Pour ma part, je me suis focalisé sur la version non payante, pour que tout un chacun puisse reproduire les opérations que je décris.

Ce document présente la mise en œuvre de Pentaho Data Integration Community Edition ([PDI-CE](#), appelée également Kettle), l'outil ETL de la suite Pentaho CE. Je me contente d'une description succincte pour deux raisons : ce type d'outil n'entre pas directement dans mon champ de compétences (qui est le data mining) ; j'en parle surtout pour préparer un prochain tutoriel dans lequel je montre le déploiement de modèles élaborés à l'aide de Knime, Sipina ou Weka via PDI-CE.

Ce tutoriel est basé sur la version stable **4.0.1** de PDI-CE (cette précision est importante !).

2 Données

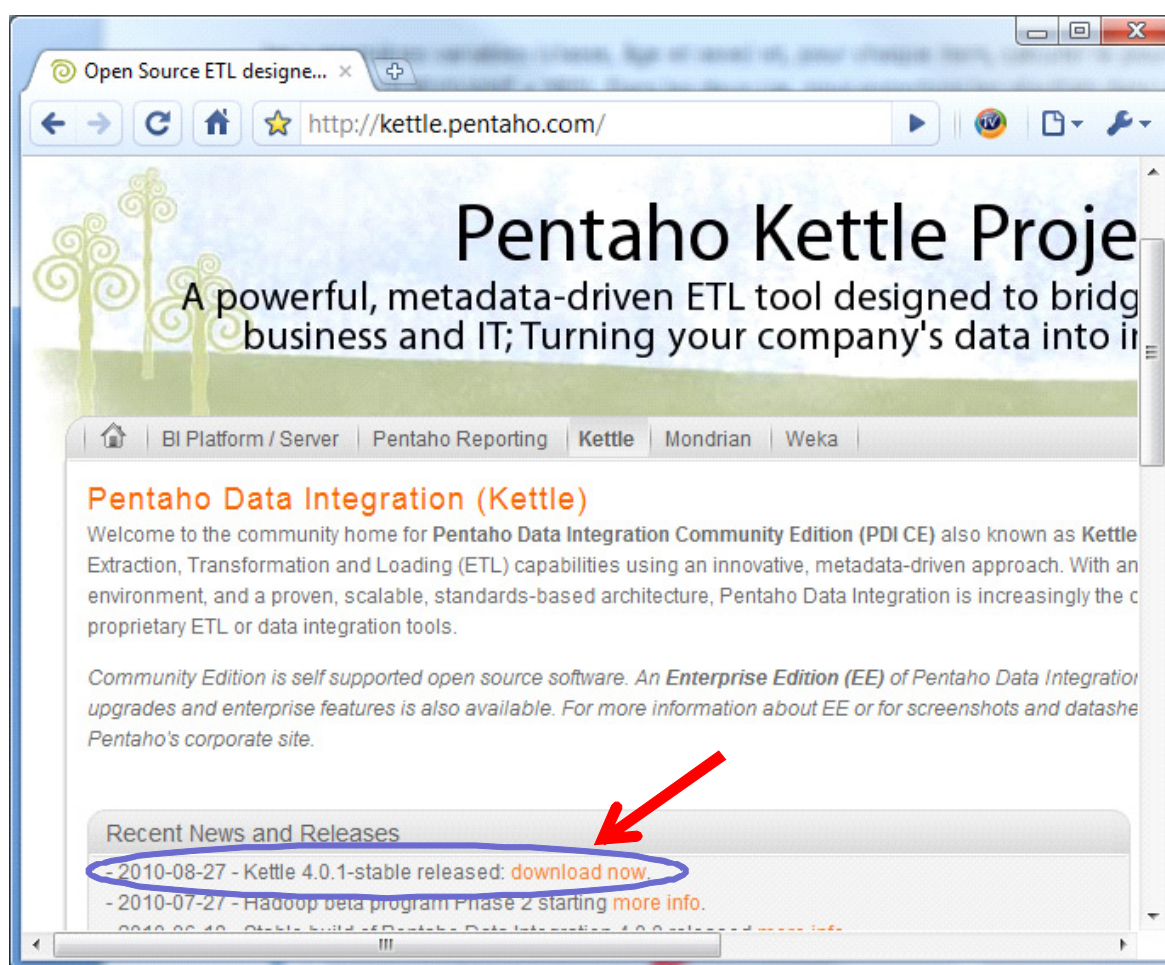
Nous utilisons une version dupliquée 32 fois ([titanic32x.csv.zip](#)) du fichier [TITANIC](#) qui recense les caractéristiques des passagers ayant participé au sinistre voyage. Sont disponibles dans la base : la classe (1^{ère} classe, 2^{nde}, 3^{ème}, membre d'équipage) ; l'âge (adulte, enfant) ; le sexe (homme, femme) et le fait d'avoir survécu ou pas (yes, no). Nous poursuivons un double objectif : (1) énumérer les différentes combinaisons (items) des 4 variables composant la base et, pour chacune d'entre elles, comptabiliser le nombre d'observations ; (2) énumérer les différentes combinaisons possibles avec

les 3 premières variables (classe, âge et sexe) et, pour chaque item, calculer le pourcentage de survivants (SURVIVANT = YES). Dans les deux cas, nous exportons les résultats dans un fichier au format Excel.

Pour pimenter la chose, nous avons dupliqué la base 32 fois, nous disposons donc de 70.432 observations. Nous aurons ainsi un meilleur aperçu des capacités de traitements de l'outil¹, rechercher des doublons dans une table de données est loin d'être une opération anodine.

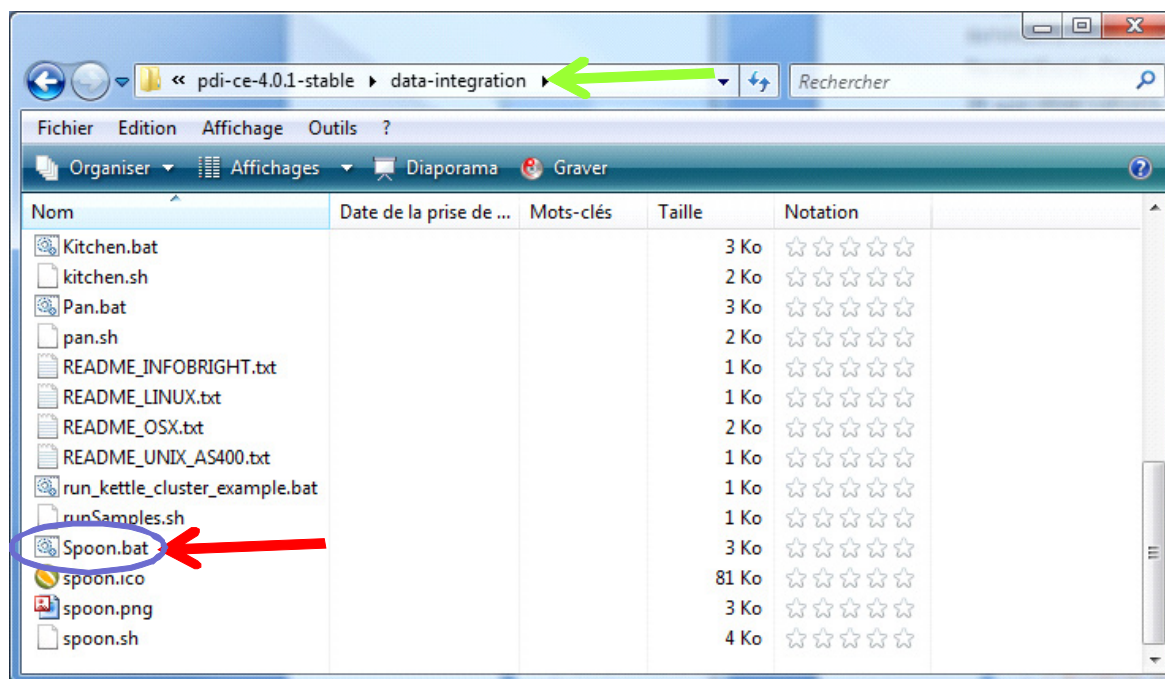
3 Chargement et installation de PDI-CE

Nous récupérons le fichier d'installation de PDI-CE 4.0.1 sur le site de Pentaho.

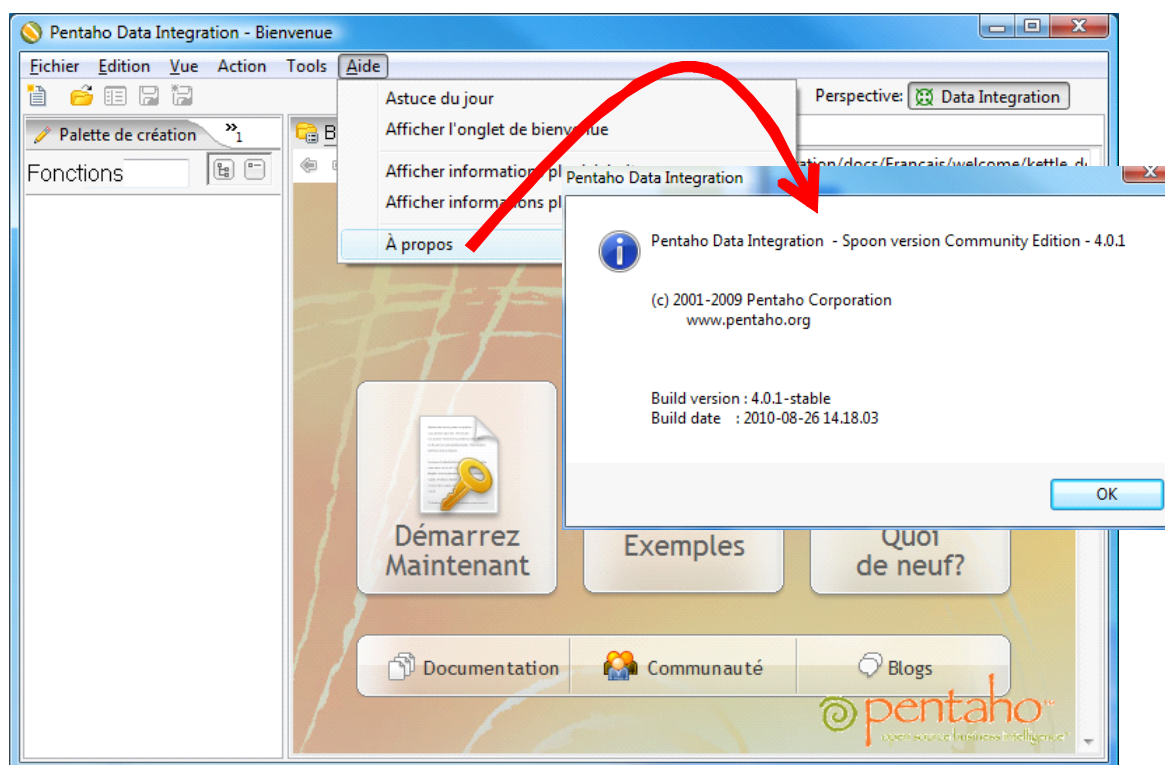


Pour installer le logiciel, il suffit de désarchiver les fichiers dans le répertoire de votre choix. SPOON.BAT permet de démarrer le logiciel.

¹ Il faudrait des tests systématiques à grande échelle, avec des configurations diversifiées (nombre de lignes, nombre de colonnes, type des colonnes, ...) pour se faire une idée plus précise des performances de l'outil.



Nous obtenons la fenêtre principale suivante. La boîte A PROPOS permet de vérifier la version réellement utilisée.

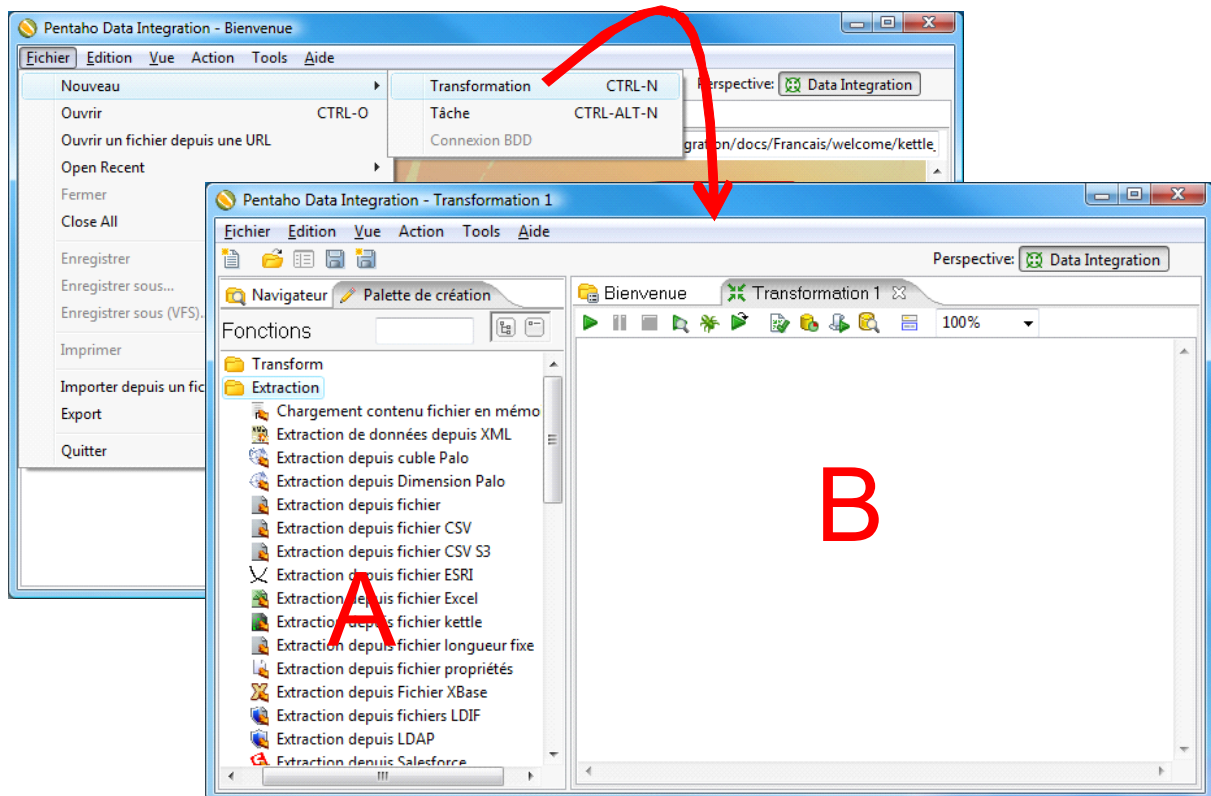


4 Création d'un projet de transformation

Pour réaliser les opérations décrites dans la section 2, nous créons un nouveau projet de transformation. Nous actionnons le menu FICHER / NOUVEAU / TRANSFORMATION.

La fenêtre principale prend un nouvel aspect. Sur la gauche, dans la palette de création (A), nous disposons des outils de manipulation de données ; sur la droite (B) un espace de travail nous

permettant de définir les séquences d'opérations sous forme de diagramme de traitements : les icônes correspondent à des opérateurs, les flèches qui le relient symbolisent les flux de données.



4.1 Enumération des valeurs et comptage

Nous disposons de 4 variables catégorielles, portant respectivement 4, 2, 2 et 2 modalités. Le nombre de combinaisons de valeurs possibles est $4 \times 2 \times 2 \times 2 = 32$. Il est évident que nous ne les aurons pas toutes, certaines n'ont pas de sens. Par exemple, un enfant ne peut pas être un membre d'équipage. Notre premier objectif dans ce tutoriel est d'énumérer les combinaisons présentes dans le fichier, puis de compter l'occurrence de chacune d'entre elles. Notre fichier comporte 70.432 observations. Nous souhaitons à la sortie obtenir un tableau avec les informations suivantes :

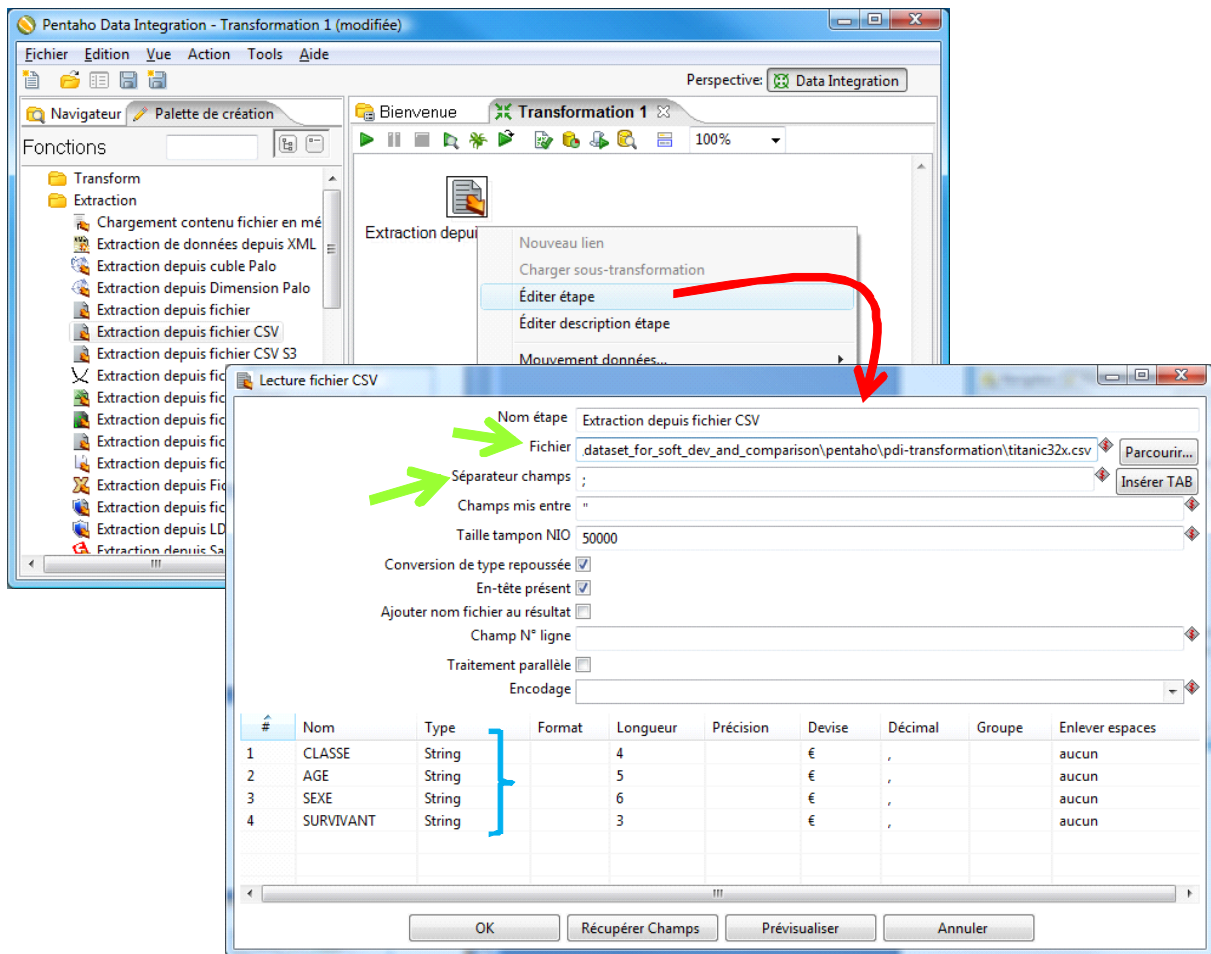
	A	B	C	D	E	F
1	CLASSE	AGE	SEXE	SURVIVANT	COUNT_SURVIVANT	
2	1ST	ADULT	FEMALE	NO	128.00	
3	1ST	ADULT	FEMALE	YES	4 480.00	
4	1ST	ADULT	MALE	NO	3 776.00	
5	1ST	ADULT	MALE	YES	1 824.00	
6	1ST	CHILD	FEMALE	YES	32.00	

Nous constatons par exemple que la combinaison (CLASSE = 1ST ; AGE = ADULT ; SEXE = FEMALE ; SURVIVANT = NO) a été observée 128 fois dans le fichier de données ; etc.

4.1.1 Accès au fichier de données

Les traitements commencent nécessairement par l'accès aux données source. Il nous faut lire le fichier « titanic32x.csv » au format CSV (fichier texte, « ; » est le caractère séparateur de colonnes).

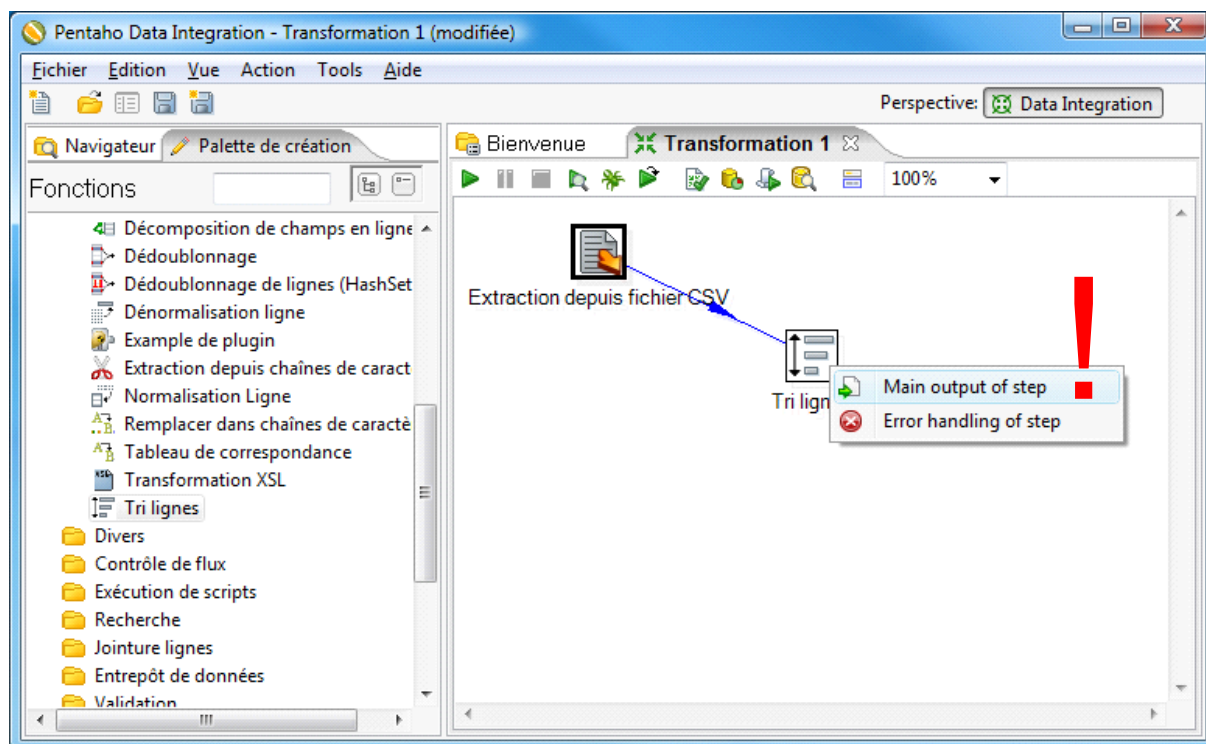
Nous introduisons le composant « **Extraction depuis le fichier CSV** » (branche Extraction) dans l'espace de travail. Nous le paramétrons en actionnant le menu contextuel « Editer étape ».



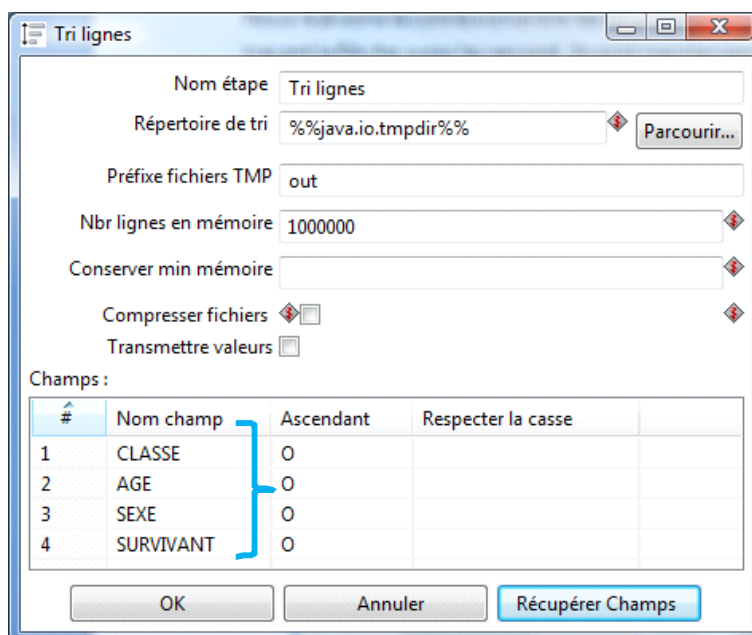
Nous spécifions le nom du fichier de données et le séparateur de champs. Pour vérifier l'intégrité du fichier, nous cliquons sur le bouton « Récupérer Champs ». L'outil reconnaît automatiquement le type de chaque colonne. Il s'appuie sur les 100 premières lignes (paramètre modifiable) du fichier pour cela. Dans notre cas, les modalités des variables catégorielles sont décrites à l'aide de chaînes de caractères (STRING).

4.1.2 Agrégation des valeurs

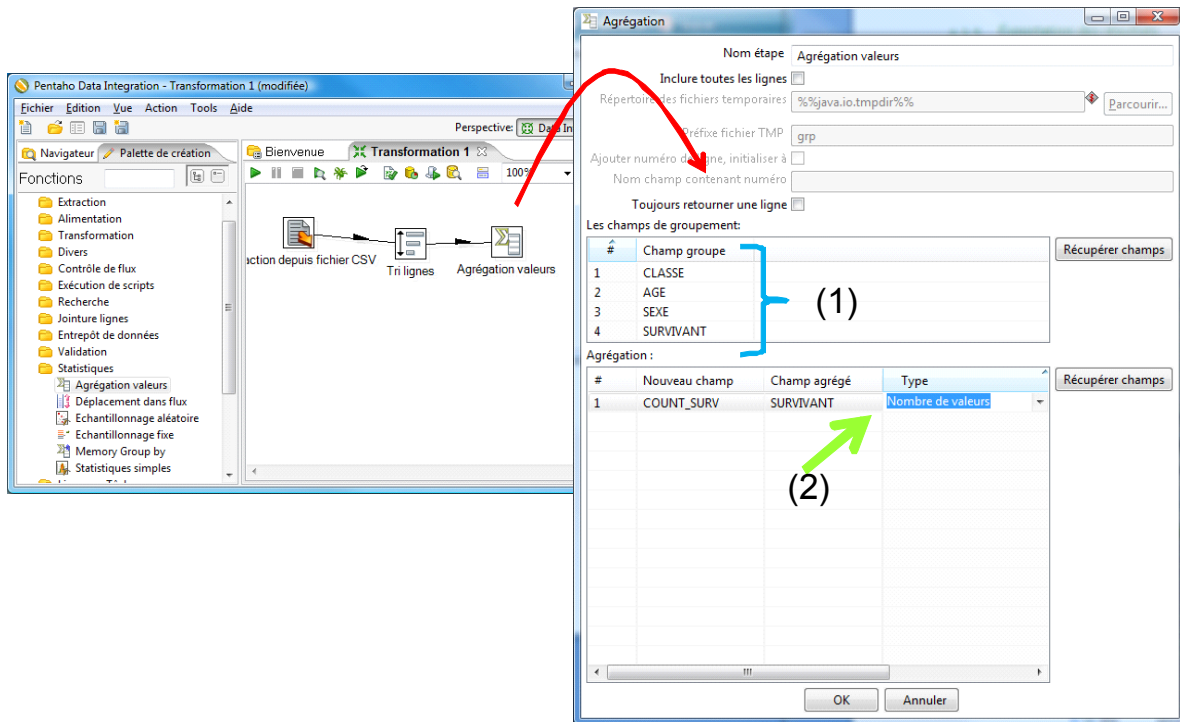
Pour comptabiliser les combinaisons des 4 variables présentes dans le fichier, puis compter leurs occurrences, il nous faut tout d'abord trier le fichier. Cela ne semble pas nécessaire au premier abord. Mais en lisant la documentation, on comprend que PDI-CE cherche les doublons en comparant l'item courant avec celui qui le précède. **Trier les données est donc une opération absolument indispensable.** Nous insérons le composant « **Tri Lignes** » (branche Transformation). Nous réalisons la connexion entre les deux outils en faisant SHIFT + Clic sur le premier icône, et en traçant la flèche jusqu'au second. Ils sont maintenant reliés. Attention, il faut confirmer la connexion en cliquant sur le menu surgissant « Main output of step ».



Nous éditons le second composant (menu contextuel « Editer étapes »). Nous cliquons sur le bouton « Récupérer les champs » pour définir les champs (tous dans notre cas) servant au tri.

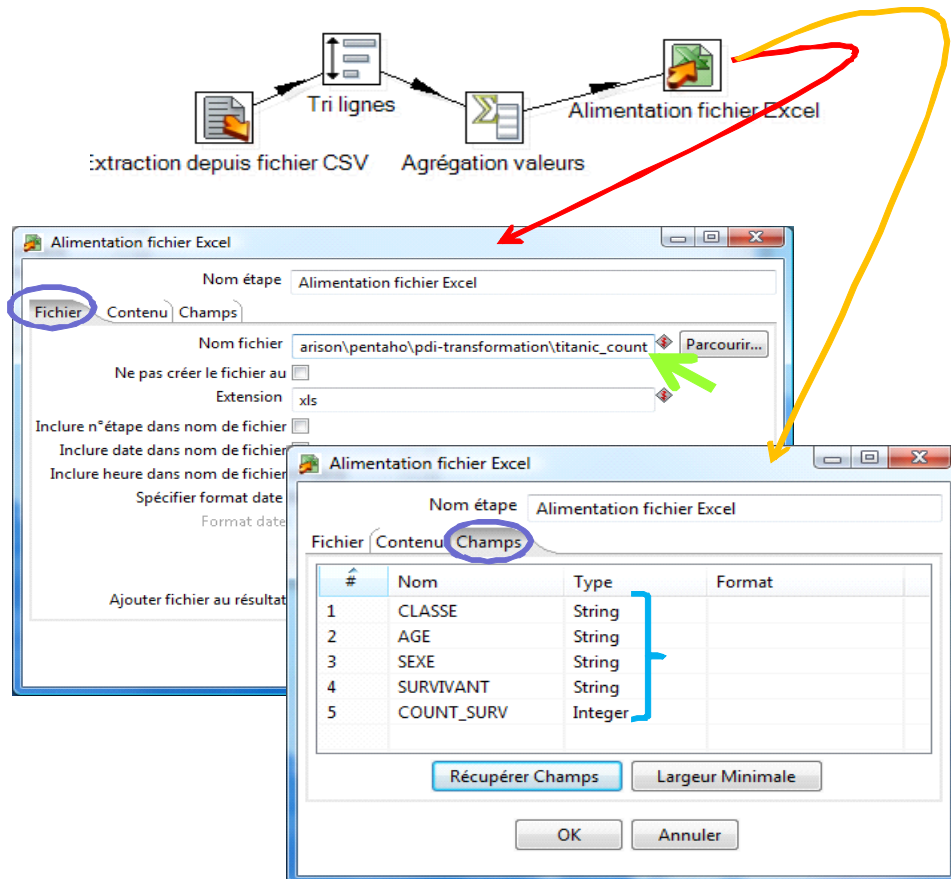


Nous pouvons passer en comptage avec le composant « **Agrégation valeurs** » (branche *Statistiques*). Après avoir réalisé la connexion avec le précédent, nous le paramétrons comme suit : (1) tous les champs doivent être pris en compte dans le regroupement ; (2) le calcul porte sur le comptage des valeurs du champ SURVIVANT (n'importe champ convenait en réalité).



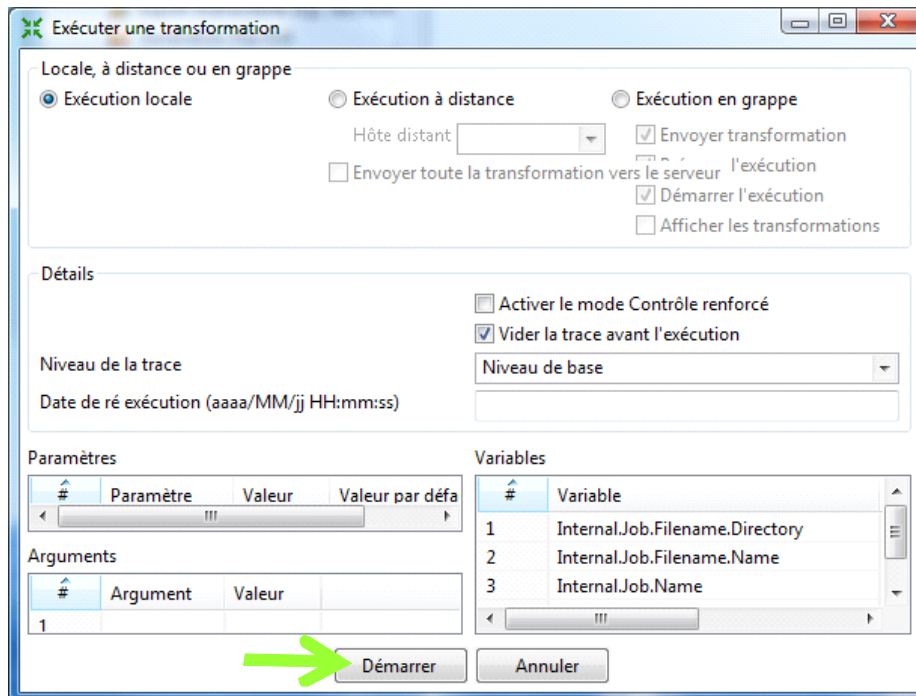
4.1.3 Exportation des résultats

Il ne reste plus qu'à exporter les résultats dans un fichier au format Excel. Nous introduisons le composant « **Alimentation fichier Excel** » (branche *Alimentation*). Après lui avoir connecté le précédent, nous le paramétrons en spécifiant le nom du fichier (onglet **Fichier**) et en précisant les champs qui doivent être exportés (onglet **Champs**).



4.1.4 Lancement des calculs

Tout est prêt maintenant. Après avoir sauvegardé le projet (Fichier / Enregistrer), nous pouvons lancer son exécution en cliquant sur le bouton ► ou en actionnant le menu Action / Exécuter (F9). Une boîte de contrôle apparaît.



Nous cliquons sur « Démarrer ». Une fenêtre relatant les statistiques d'exécution apparaît dans la partie basse de l'espace de travail.

The 'Résultats exécution' window displays the following table:

#	Nom étape	N°Copie	Lignes lues	Lignes écrites
1	Extraction depuis fichier CSV	0	0	70432
2	Tri lignes	0	70432	70432
3	Agrégation valeurs	0	70432	24
4	Alimentation fichier Excel	0	24	24

The 'Lignes lues' and 'Lignes écrites' columns for step 3 are circled in blue, and a red exclamation mark is visible in the 'N°Copie' column for step 3.

Le fichier contient 70.432 lignes. Après agrégation, nous observons 24 combinaisons de valeurs distinctes. Nous consultons le fichier « titanic_count.xls » pour obtenir le détail des résultats.

	A	B	C	D	E
1	CLASSE	AGE	SEXE	SURVIVANT	COUNT_SURV
2	1ST	ADULT	FEMALE	NO	128.00
3	1ST	ADULT	FEMALE	YES	4 480.00
4	1ST	ADULT	MALE	NO	3 776.00
5	1ST	ADULT	MALE	YES	1 824.00
6	1ST	CHILD	FEMALE	YES	32.00
7	1ST	CHILD	MALE	YES	160.00
8	2ND	ADULT	FEMALE	NO	416.00
9	2ND	ADULT	FEMALE	YES	2 560.00
10	2ND	ADULT	MALE	NO	4 928.00
11	2ND	ADULT	MALE	YES	448.00
12	2ND	CHILD	FEMALE	YES	416.00
13	2ND	CHILD	MALE	YES	352.00
14	3RD	ADULT	FEMALE	NO	2 848.00
15	3RD	ADULT	FEMALE	YES	2 432.00
16	3RD	ADULT	MALE	NO	12 384.00
17	3RD	ADULT	MALE	YES	2 400.00
18	3RD	CHILD	FEMALE	NO	544.00
19	3RD	CHILD	FEMALE	YES	448.00
20	3RD	CHILD	MALE	NO	1 120.00
21	3RD	CHILD	MALE	YES	416.00
22	CREW	ADULT	FEMALE	NO	96.00
23	CREW	ADULT	FEMALE	YES	640.00
24	CREW	ADULT	MALE	NO	21 440.00
25	CREW	ADULT	MALE	YES	6 144.00
26					

Lorsque nous effectuons la somme des valeurs de COUNT_SURV, nous retrouvons 70.432, soit le nombre d'observations dans la base.

C'est absolument charmant. Surtout si l'on considère le temps de calcul, très rapide au regard des opérations réalisées (chargement des données, tri, comptage, écriture du fichier de sortie).

4.2 Enumération des valeurs et calcul des fréquences

Dans cette deuxième partie, nous souhaitons comptabiliser la proportion de SURVIVANT = YES pour chaque combinaison des variables CLASSE, AGE et SEXE. Reprenons le résultat ci-dessus. Nous observons qu'il y a $(128 + 4480) = 4608$ individus (CLASSE = 1ST, AGE = ADULT, SEXE = FEMALE); 97.22% $(128 / 4608)$ d'entre eux ont survécu au naufrage (SURVIVANT = YES). Nous obtiendrons un tableau ressemblant à ceci :

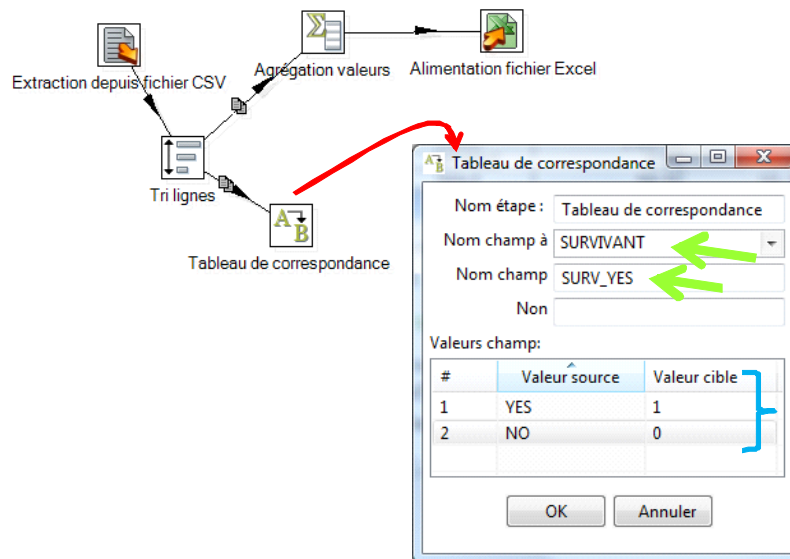
	A	B	C	D	E	F
1	CLASSE	AGE	SEXE	COUNT_SURVIVANT	FREQ_SURV_YES	
2	1ST	ADULT	FEMALE	4 608.00	97.22%	
3	1ST	ADULT	MALE	5 600.00	32.57%	
4	1ST	CHILD	FEMALE	32.00	100.00%	
5	1ST	CHILD	MALE	160.00	100.00%	

Nous observons dans les deux dernières colonnes : le nombre d'observations pour chaque item et la proportion de SURVIVANT = YES.

4.2.1 Recodage de la variable SURVIVANT

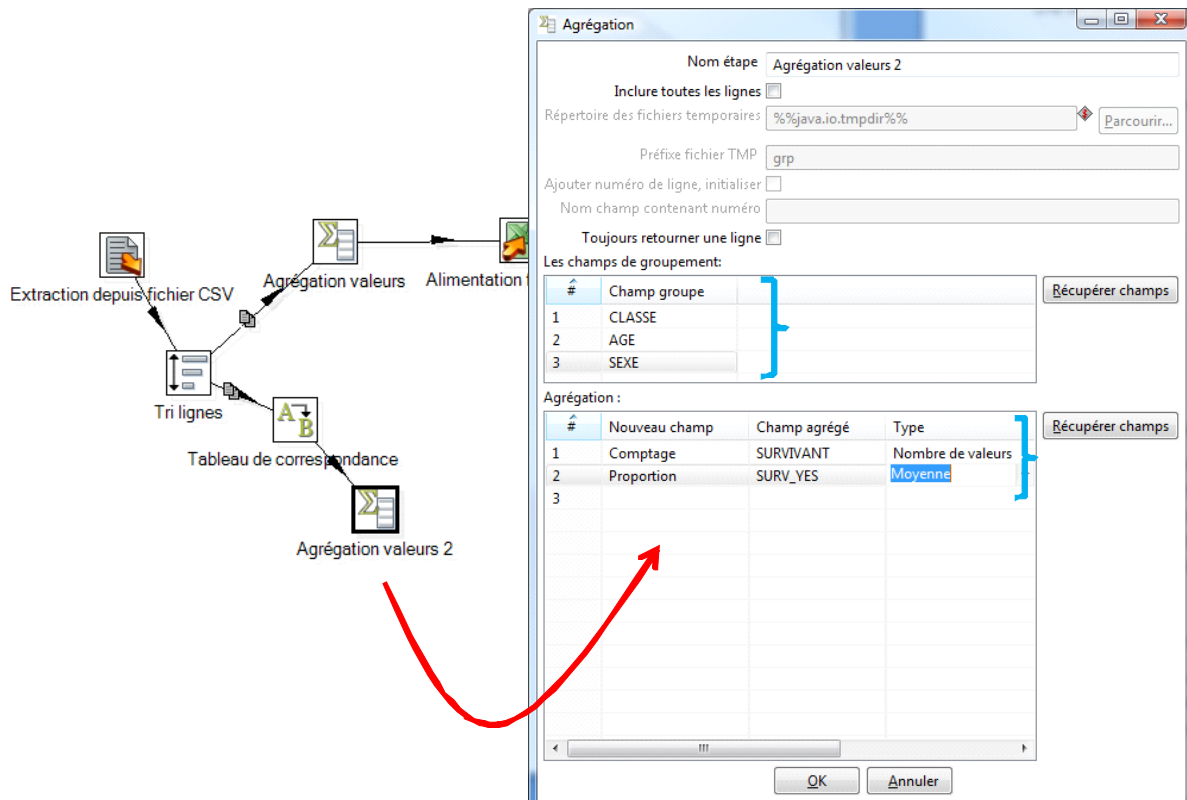
Pour calculer les proportions, nous devons recoder la variable SURVIVANT en variable binaire SURV_YES : 1 lorsque SURVIVANT = YES, 0 autrement. Ainsi, en calculant la moyenne de la variable recodée, nous obtenons naturellement la proportion de YES.

Pour cela, nous utilisons le composant « **Tableau de correspondance** » (branche *Transformation*). Nous lui connectons le tableau déjà trié (ce n'est pas indispensable, mais cela évite d'avoir à refaire une seconde fois le tri par la suite), nous le paramétrons de la manière suivante.



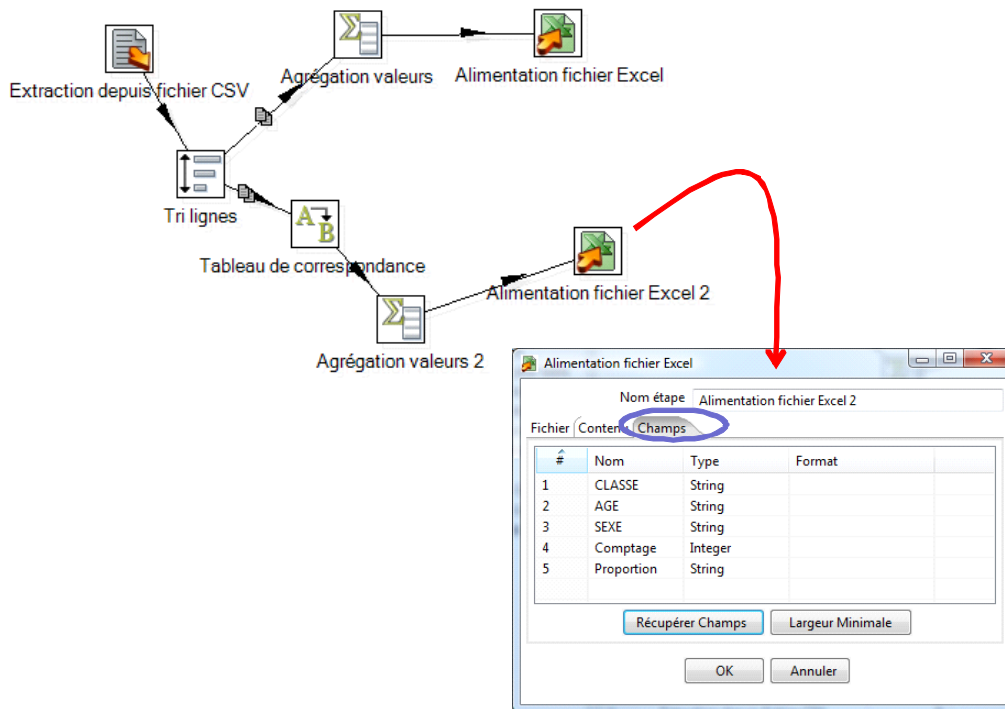
4.2.2 Agrégation des valeurs

Nous pouvons définir les calculs via le composant « **Agrégation valeurs** » (branche *Statistiques*). Deux opérations sont nécessaires pour chaque combinaison de CLASSE, AGE et SEXE (champs de regroupement) : comptage des observations (Comptage = Nombre de valeurs de SURVIVANT) ; calcul de la proportion de SURVIVANT = YES (Proportion = Moyenne sur SURV_YES). Nous adoptons le paramétrage suivant :



4.2.3 Exportation des résultats

Il ne nous reste plus qu'à exporter le résultat à l'aide du composant « Alimentation fichier Excel ». Nous créons le fichier « **titanic_freq.xls** ». Nous y intégrons les champs CLASSE, AGE et SEXE, ainsi que les champs calculés « Comptage » et « Proportion ».



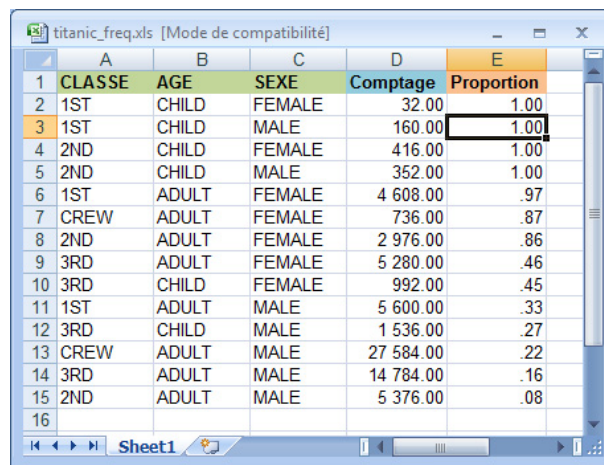
4.2.4 Lancement des calculs

Nous actionnons le bouton ▶ pour lancer les calculs. Dans la fenêtre de suivi, PDI-CE nous indique que 14 lignes de valeurs ont été générées dans le fichier de sortie.

The screenshot shows the Pentaho Data Integration interface. The job flow is visible in the main window. Below it, the 'Résultats exécution' window is open, showing a table of execution results. A green arrow points to the value '14' in the 'Lignes écrites' column for the final 'Alimentation fichier Excel 2' step.

#	Nom étape	N° Copie	Lignes lues	Lignes écrites
1	Extraction depuis fichier CSV	0	0	70432
2	Tri lignes	0	70432	140864
3	Agréation valeurs	0	70432	24
4	Alimentation fichier Excel	0	24	24
5	Tableau de correspondance	0	70432	70432
6	Agréation valeurs 2	0	70432	14
7	Alimentation fichier Excel 2	0	14	14

En ouvrant le fichier dans Excel, nous obtenons le tableau de résultats désiré. Nous l'avons trié selon une proportion décroissante de SURVIVANT = YES. Nous aurions pu également spécifier cette opération directement dans PDI-CE.



	A	B	C	D	E
1	CLASSE	AGE	SEXE	Comptage	Proportion
2	1ST	CHILD	FEMALE	32.00	1.00
3	1ST	CHILD	MALE	160.00	1.00
4	2ND	CHILD	FEMALE	416.00	1.00
5	2ND	CHILD	MALE	352.00	1.00
6	1ST	ADULT	FEMALE	4 608.00	.97
7	CREW	ADULT	FEMALE	736.00	.87
8	2ND	ADULT	FEMALE	2 976.00	.86
9	3RD	ADULT	FEMALE	5 280.00	.46
10	3RD	CHILD	FEMALE	992.00	.45
11	1ST	ADULT	MALE	5 600.00	.33
12	3RD	CHILD	MALE	1 536.00	.27
13	CREW	ADULT	MALE	27 584.00	.22
14	3RD	ADULT	MALE	14 784.00	.16
15	2ND	ADULT	MALE	5 376.00	.08
16					

5 Conclusion

Ce document donne un aperçu très succinct des capacités de PDI-CE en matière de management des données. Il est possible de définir des travaux de chargement, de manipulation et de nettoyage de données, sans avoir à écrire une seule ligne de programme.

La question de la volumétrie reste cependant posée. Je ne l'ai pas vraiment explorée. J'y reviendrai vraisemblablement dans un prochain didacticiel. Cet aspect est essentiel pour un outil ETL.