

# 1 Topic

## Short description of the Pentaho Data Integration Community Edition (Kettle).

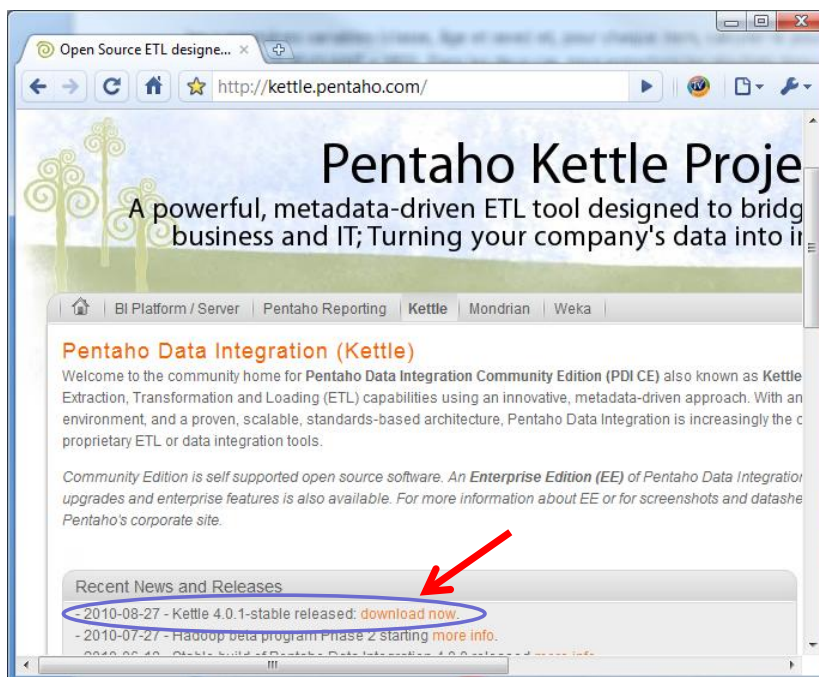
The Pentaho BI Suite is an open source Business Intelligence suite with integrated reporting, dashboard, data mining, workflow and ETL capabilities (<http://en.wikipedia.org/wiki/Pentaho>)<sup>1</sup>. In this tutorial, we talk about the Pentaho BI Suite Community Edition (CE) which is freely downloadable. More precisely, we present the Pentaho Data Integration (PDI-CE)<sup>2</sup>, called also Kettle<sup>3</sup>. We show briefly how to load a dataset and perform a simplistic data analysis. The main goal of this tutorial is to introduce a next one focused on the deployment of the models designed with Knime, Sipina or Weka by using PDI-CE.

This document is based on the **4.0.1** stable version of PDI-CE.

## 2 Dataset

We have duplicated the **TITANIC** dataset 32 times in this tutorial ([titanic32x.csv.zip](#)). We have 4 variables: CLASSE (CLASS), AGE, SEXE (SEX), SURVIVANT (SURVIVED). We have duplicated the rows in order to evaluate the ability to handle a large dataset (70,432 rows and 4 columns is not really a large dataset, but the initial database was really too small).

We have two goals: (1) enumerate different combinations (items) of the 4 variables, and for each of them, count the number of observations; (2) enumerate the possible combinations for the first 3 variables (class, age and gender) and, for each item, calculate the percentage of survivors (SURVIVANT = YES). We export the results into a file in the Excel format.



## 3 Loading and installing PDI-CE

We load the setup file on the Pentaho Website (PDI-CE 4.0.1)<sup>4</sup>.

To install the software, we simply expand the archive file into a directory. We launch the tool by clicking on the

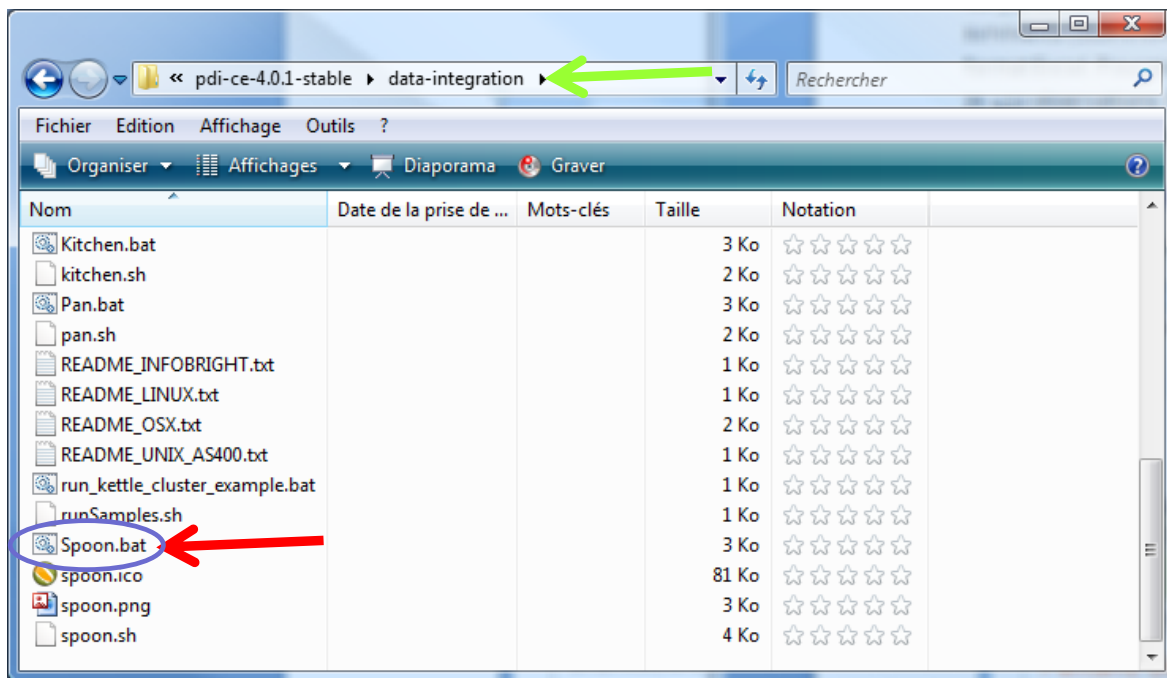
<sup>1</sup> <http://www.pentaho.com/>

<sup>2</sup> <http://community.pentaho.com/>

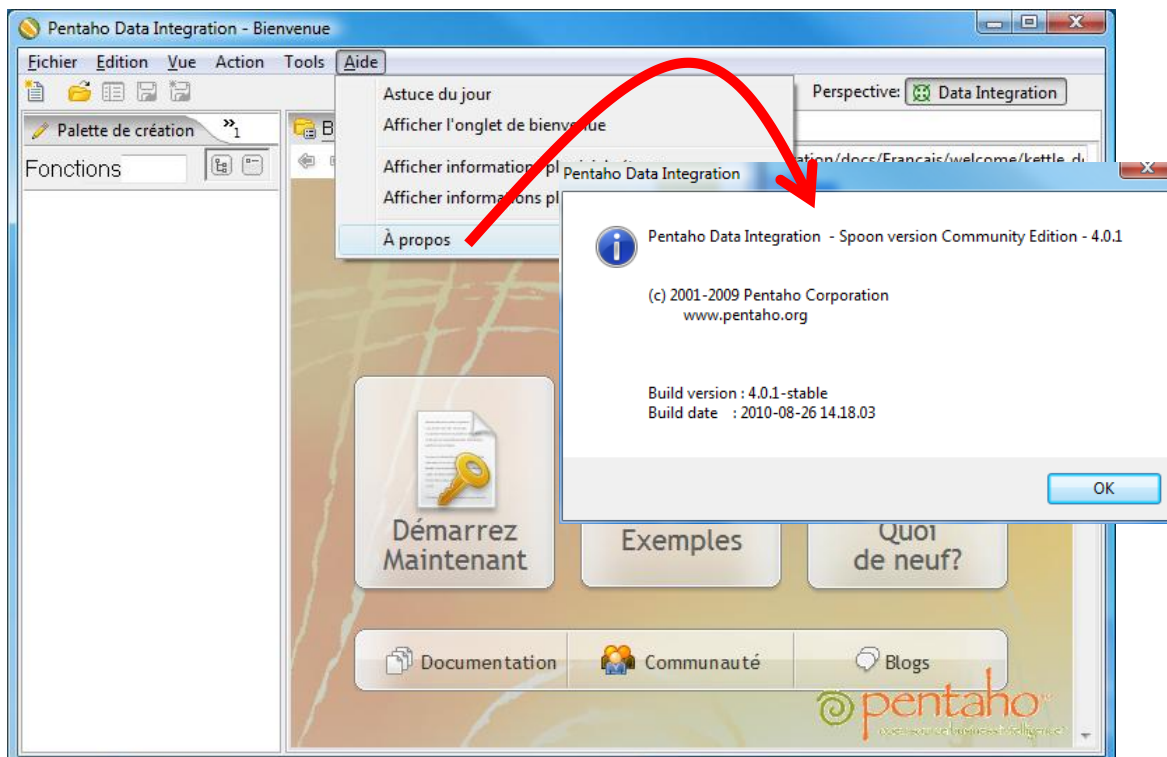
<sup>3</sup> <http://kettle.pentaho.com/>

<sup>4</sup> We have written the French version of this tutorial in September 2010. The current version of PDI-CE is 4.2 (2011/09/12). But I hope that the descriptions remain valid.

**SPOON.BAT** file.

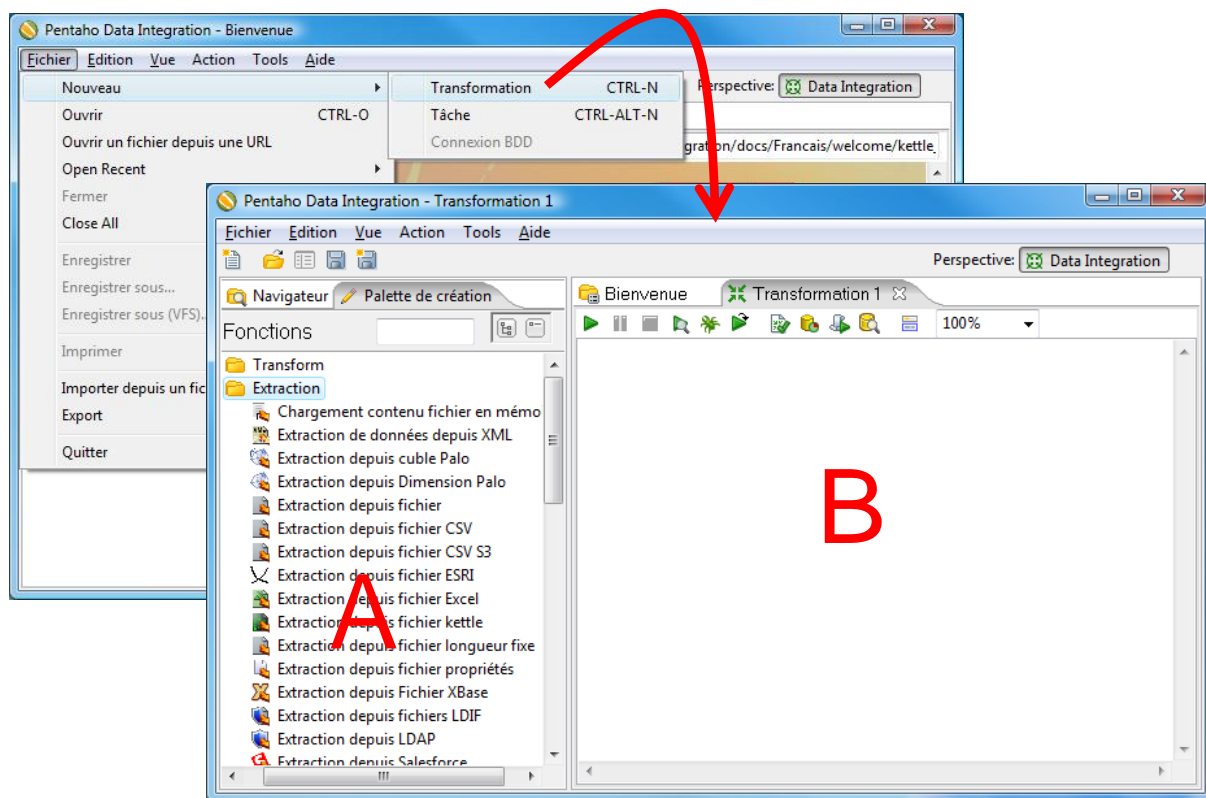


We obtain the following main window. The about box enables to check the version used.



## 4 Creating a project

To create a new project, we click on the FICHIER / NOUVEAU / TRANSFORMATION menu. The main window is reorganized. At the left, we have the CREATION PALETTE. At the right, we have the workspace. It enables us to define the data transformation workflow.



#### 4.1 Enumerating and counting itemsets

We have 4 variables with respectively 4, 2, 2 and 2 values. Then, we should obtain 32 ( $4 \times 2 \times 2 \times 2$ ) itemsets of length 4. But some of items are not consistent. For instance, a child cannot be a crew member. So, the number of observed itemsets is lower than 32.

At the end of the process, we want to obtain a result table similar to the following one.

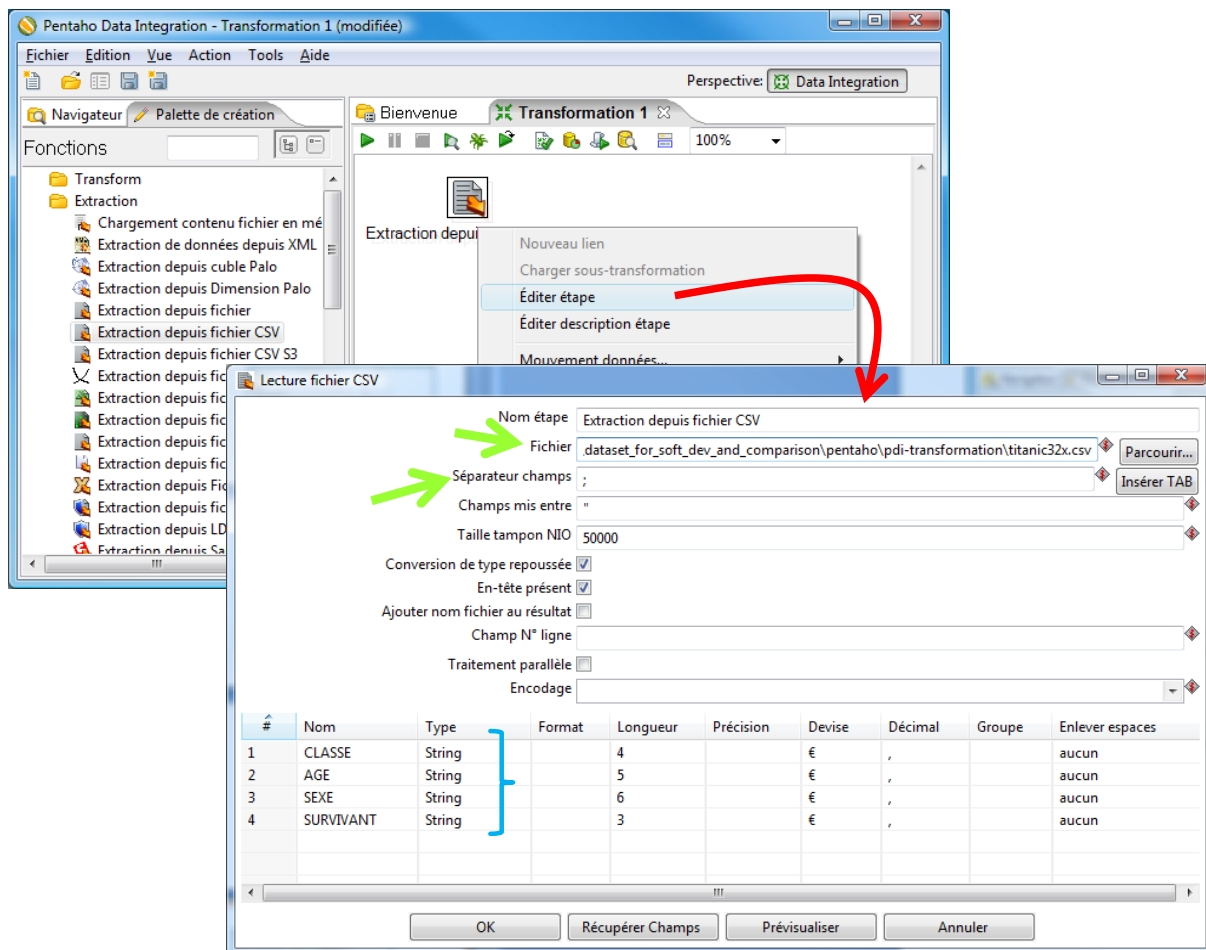
	A	B	C	D	E	F
1	CLASSE	AGE	SEXE	SURVIVANT	COUNT_SURVIVANT	
2	1ST	ADULT	FEMALE	NO	128.00	
3	1ST	ADULT	FEMALE	YES	4 480.00	
4	1ST	ADULT	MALE	NO	3 776.00	
5	1ST	ADULT	MALE	YES	1 824.00	
6	1ST	CHILD	FEMALE	YES	32.00	

We observe for instance that 128 rows (individuals) corresponds to the itemset (CLASSE = 1ST; AGE = ADULT; SEXE = FEMALE; SURVIVANT = NO) in the database; etc.

##### 4.1.1 Reading the data file

First, we must read the data file (CSV file format, “;” is the column separator). We add the “**Extraction depuis le fichier CSV**” component into the workspace. We specify the settings by clicking on the “Editer Etape” contextual menu<sup>5</sup>.

<sup>5</sup> I use the French version of PDI-CE in this tutorial, but I think the reader can easily transpose the tutorial to other language versions.

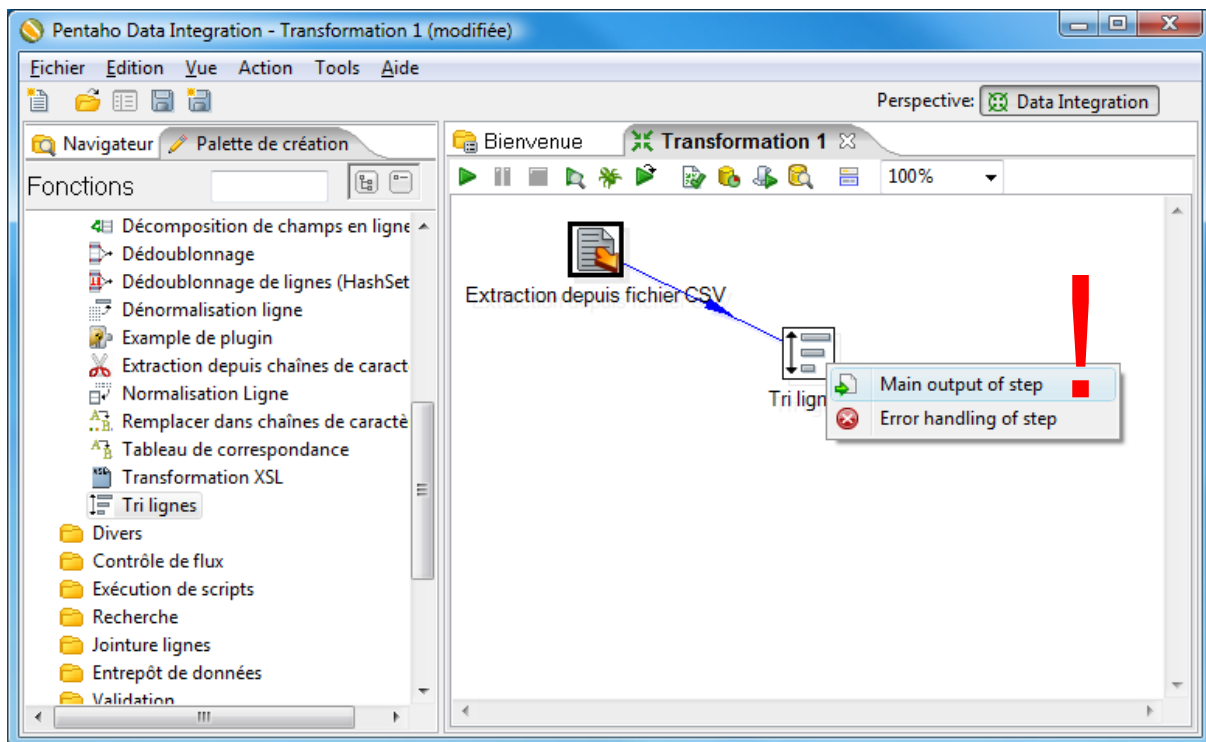


To check of the data reading, we click on the “**Récupérer Champs**” button. The tool determines automatically the column data type. It uses the 100 first rows for that. We can modify this setting. In our dataset, the categorical variables are defined as STRING.

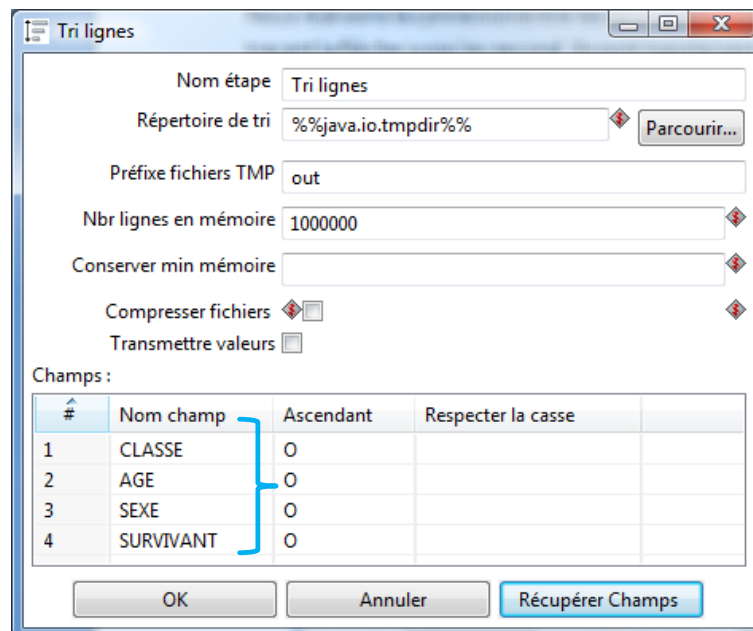
#### 4.1.2 Calculating the frequency of itemsets

To count the combinations of 4 variables present in the database, then to count their frequencies, we must first sort the file. This does not seem required at first sight. But by reading the documentation, we understand that PDI-CE searches duplicates by comparing the item current with the preceding. Thus, **sorting the data is therefore a very important preliminary operation**.

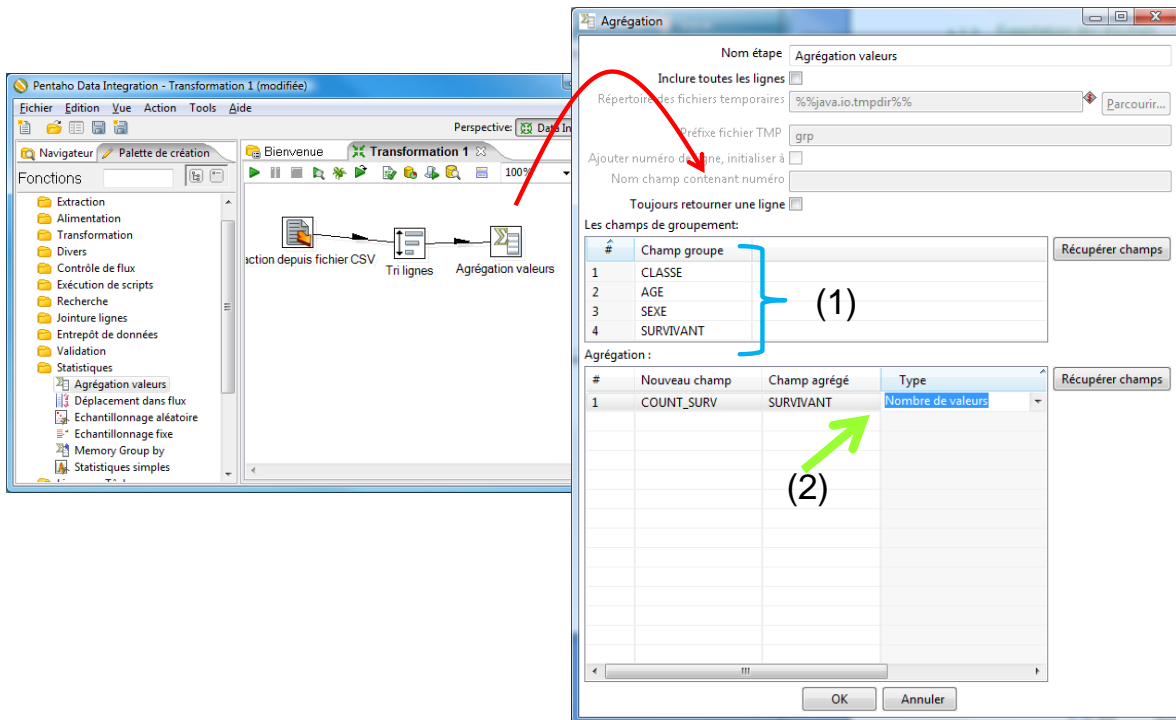
We add the “**Tri Lignes**” component into the workspace. We make the connection between the two tools by using SHIFT + CLICK on the first node. Then we draw the arrow to the second node. We must confirm the connection by clicking on the “Main output of step” contextual menu.



We edit the second component ("Edit Etapes" contextual menu). We click on the button "Récupérer champs" to define the fields used for sorting.

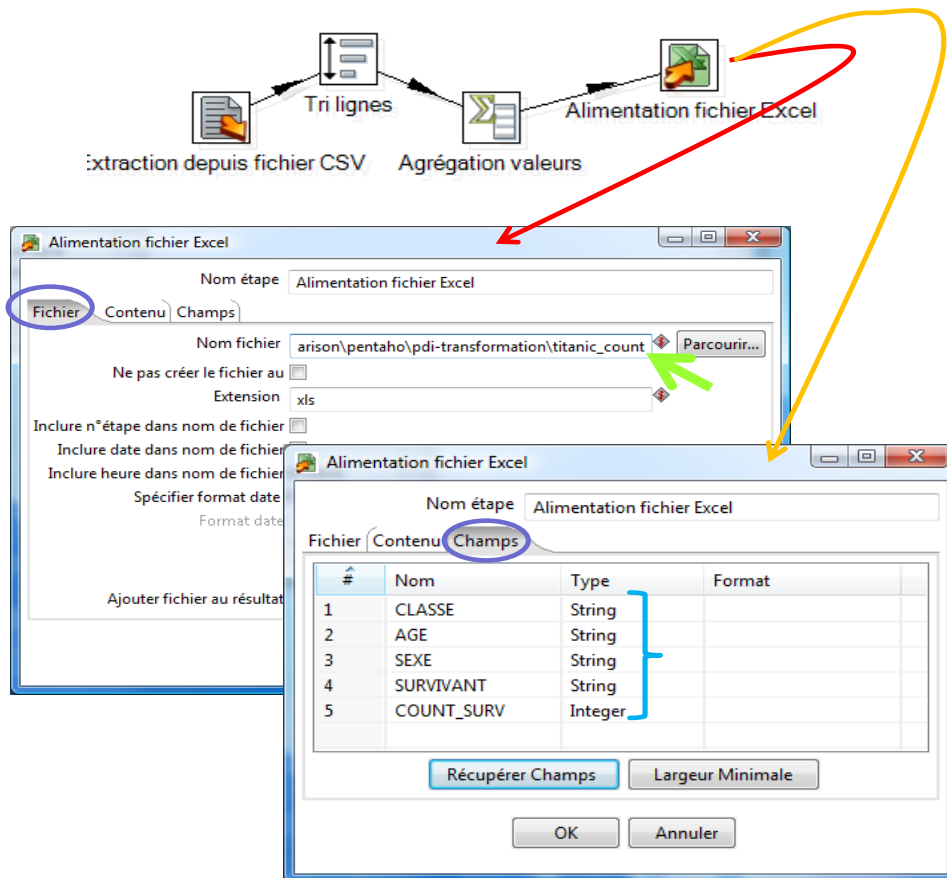


We perform the counting using the "Agrégation valeurs" component. After we connect the preceding node to this last one, we set the following parameters: (1) we use all the fields for the itemsets mining; (2) we use SURVIVANT for the counting (any field would be appropriate actually).



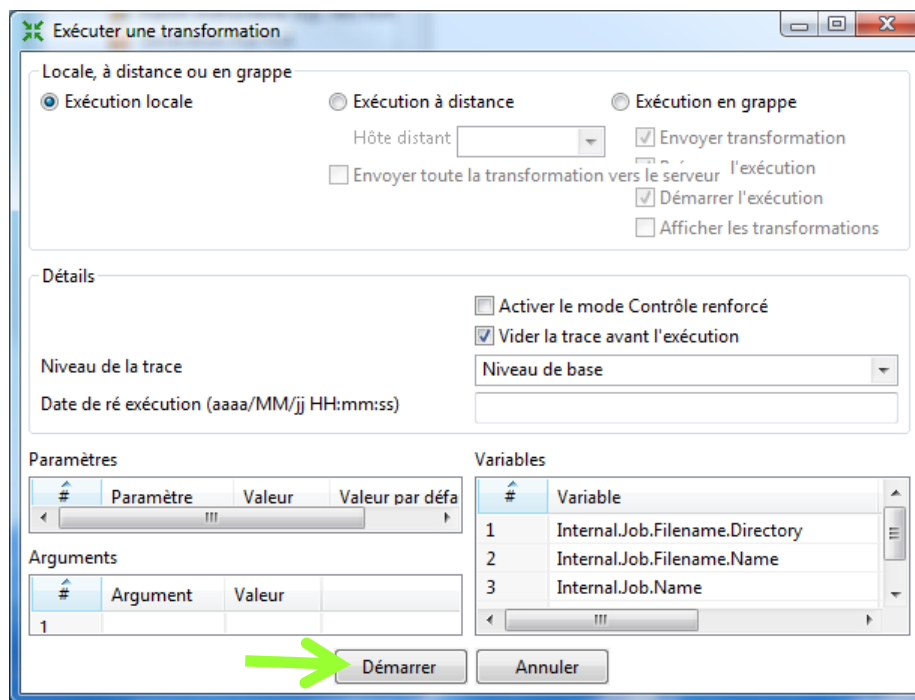
#### 4.1.3 Exporting the results

Last, we want to export the results in a file with the XLS format. We add the “**Alimentation Fichier Excel**” into the Workspace. We make the connection with the previous node. We set the appropriate parameters i.e. the data file name (**Fichier** tab) and the name of the fields to export (**Champs** tab).



#### 4.1.4 Launching the calculations

We are now ready to launch the calculations. We save the project. Then we click on the ► button or we click on the Action / Executer [F9] menu. A dialog box appears.



We click on the “Démarrer” button. A window describing the execution statistics appears in the lower part of the workspace.

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

The data file contains 70.432 rows. We obtain 24 distinct itemsets. The results are exported into the “titanic\_count.xls” file.

	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					

The sum of the COUNT\_SURV column is equal to the number of instances into the database. I noted that the calculation is really fast on our database.

## 4.2 Counting frequencies

In this section, we want to calculate the proportion of SURVIVANT = YES for each combination of CLASSE, AGE and SEXE. Let us take the results above: 4608 (128 + 4480) instances correspond to the characteristics (CLASSE = 1ST, AGE = ADULT, SEXE = FEMALE); 128 had survived (CLASSE = 1ST, AGE = ADULT, SEXE = FEMALE, SURVIVANT = YES); we obtain the ratio 97.22% (128 / 4608). The results are summarized in a table as follows.

	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%	

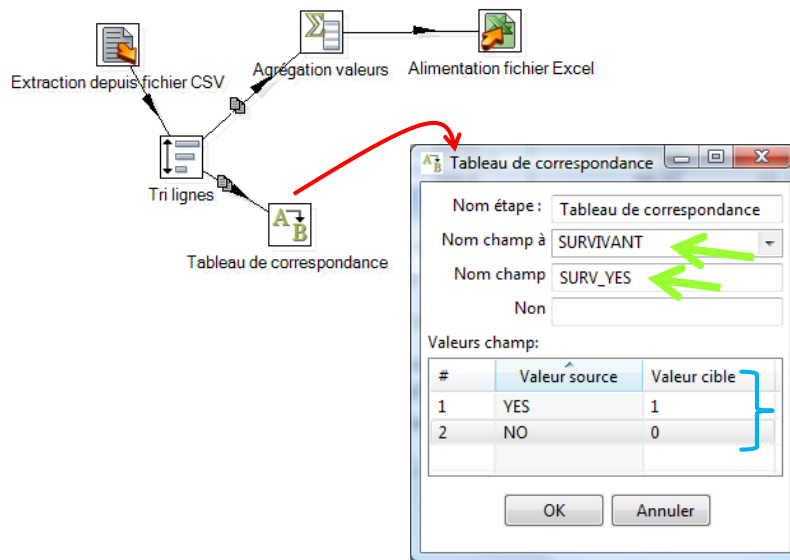
Into the last two columns, we have the number of instances covered by itemset (COUNT\_SURVIVANT), and the proportion of survivors (FREQ\_SURV\_YES).

### 4.2.1 Coding the SURVIVANT field

To calculate the proportions, we must recode the survivor variable as a binary one (SURV\_YES): 1 when SURVIVANT = 1, 0 otherwise. Thus, when we compute the mean of this new column, we obtain the proportion of SURVIVANT=YES.

To do this, we use the "Tableau de correspondance" component. We connect the sorted array (this avoids to redo twice the sorting operation afterwards). We set the following settings.

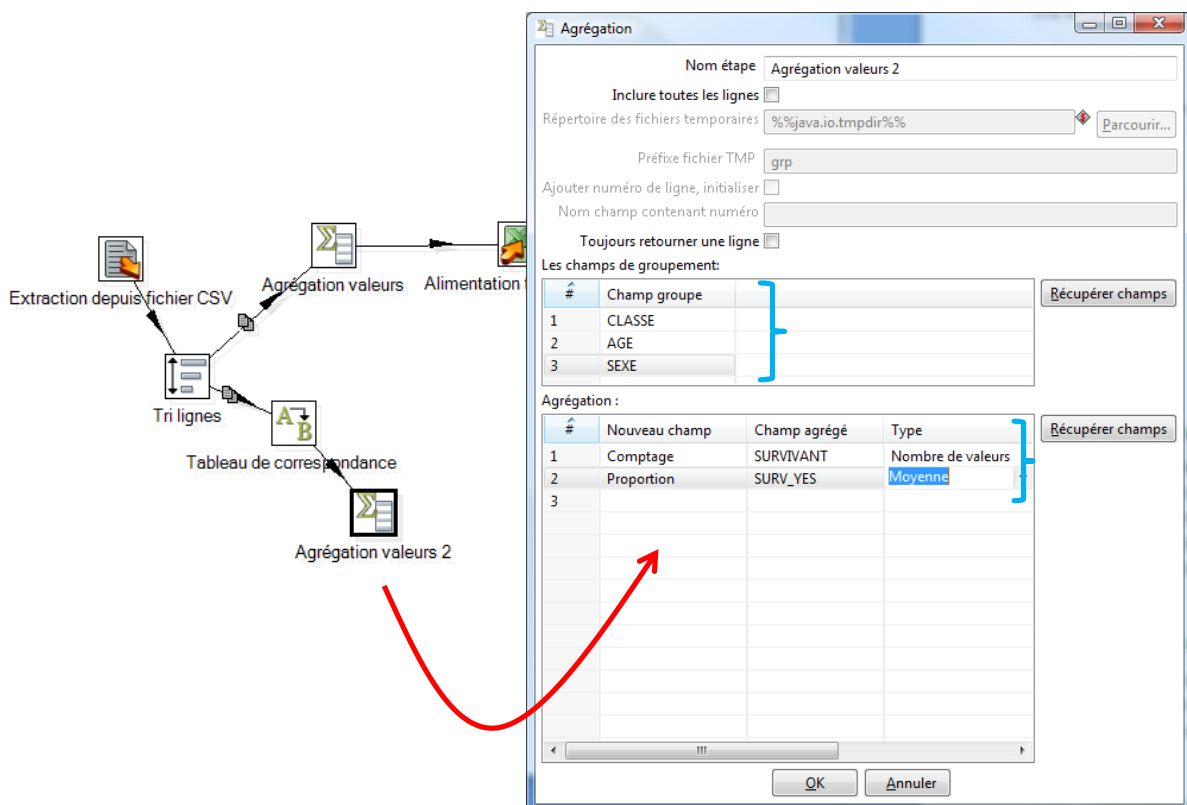




### 4.2.2 Aggregating the rows

We specify the analysis with the “**Agréation valeurs**” component. Two calculations are needed for each occurrence of the itemsets incorporating CLASSE, AGE and SEXE: (1) counting the corresponding instances, (2) computing the frequencies of SURVIVANT = YES.

We set the following parameters into the dialog settings.



**Agréation**

Nom étape: Agrégation valeurs 2

Inclure toutes les lignes:

Répertoire des fichiers temporaires: %%java.io.tmpdir%%

Préfixe fichier TMP: grp

Ajouter numéro de ligne, initialiser:

Nom champ contenant numéro:

Toujours retourner une ligne:

Les champs de groupement:

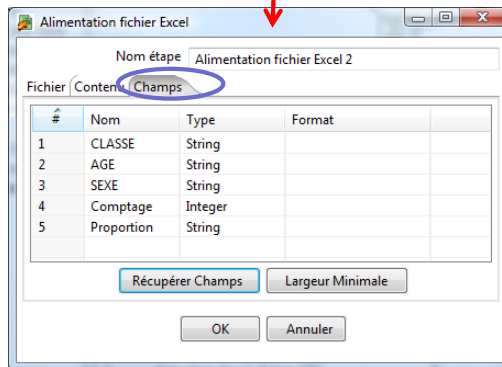
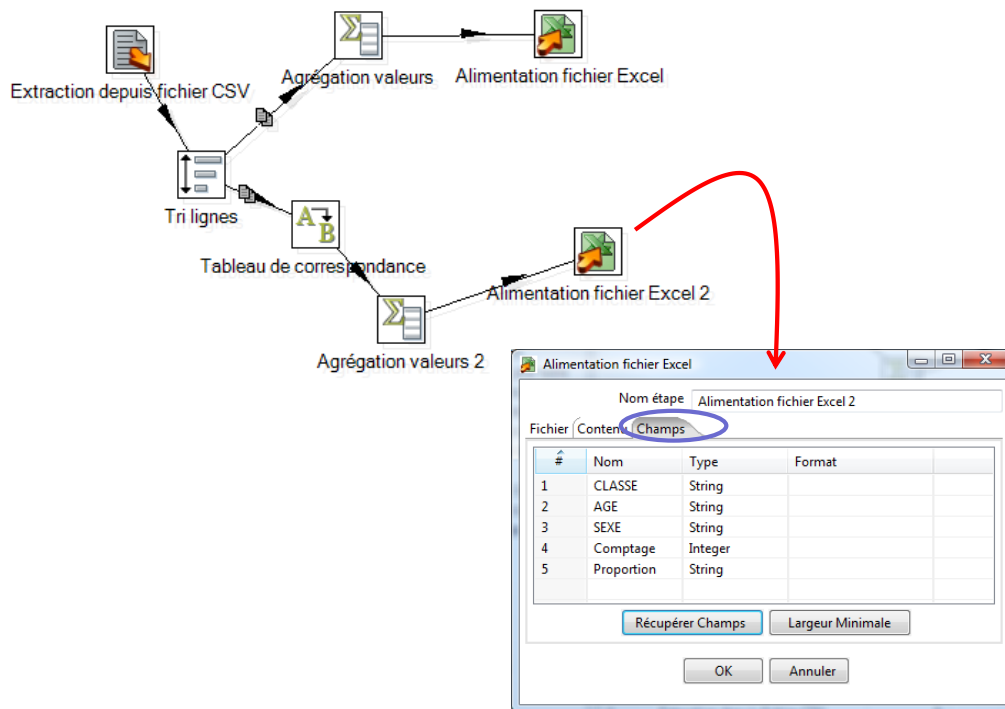
#	Champ groupe
1	CLASSE
2	AGE
3	SEXE

Agrégation:

#	Nouveau champ	Champ agrégé	Type
1	Comptage	SURVIVANT	Nombre de valeurs
2	Proportion	SURV_YES	Moyenne
3			

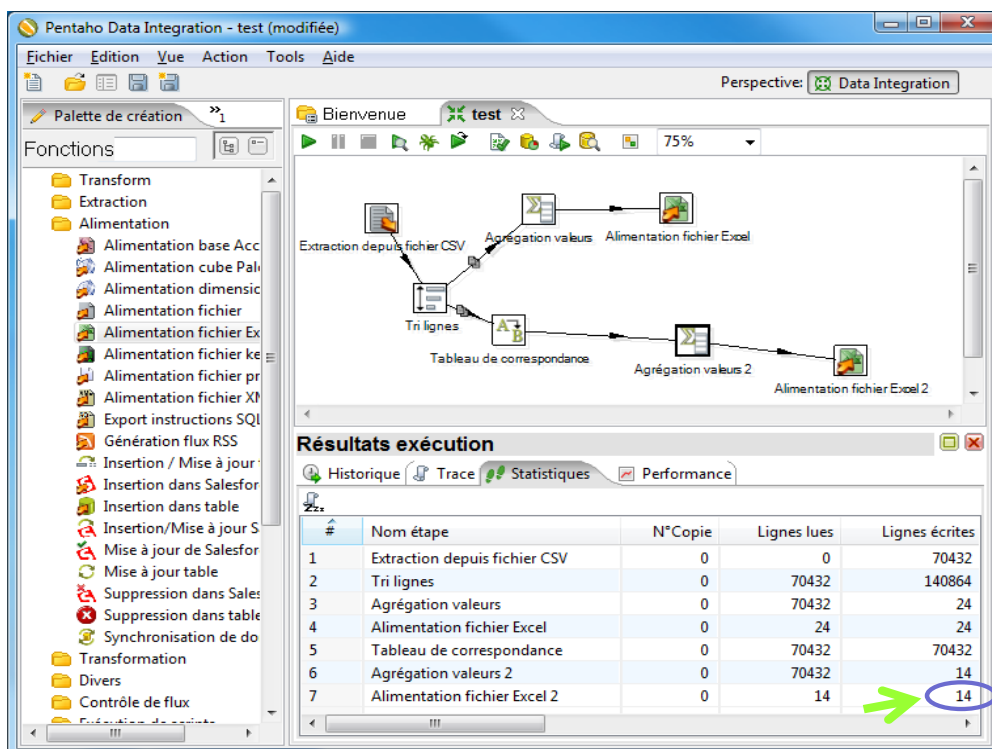
### 4.2.3 Exporting the results

We export the results with the "Alimentation fichier Excel" component. We create the "titanic\_freq.xls" Excel file. We insert the CLASSE, AGE and SEXE field; we set also the calculated fields "Comptage" (counting) and "Proportion".

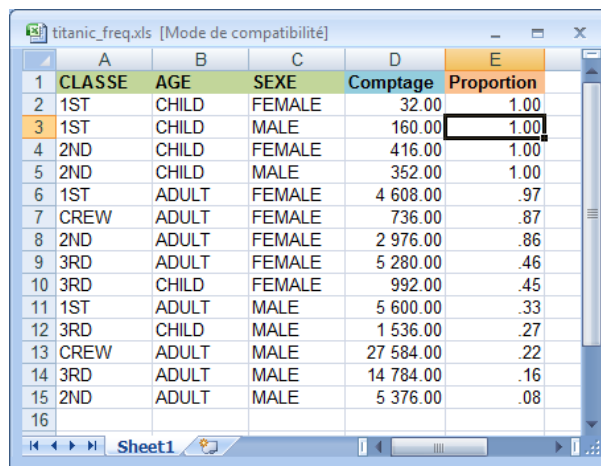


### 4.2.4 Launching the calculations

We click on the ► button to launch calculations. Into the monitoring window, PDI-CE shows that 14 rows have been generated into the output file.



When we open the output file, we get the table below. We have manually sorted the results according the proportion column (in a decreasing order). We could also define this operation directly in PDI-CE.



The screenshot shows an Excel spreadsheet titled 'titanic\_freq.xls [Mode de compatibilité]'. The data is presented in a table with the following columns: CLASSE, AGE, SEXE, Comptage, and Proportion. The rows are sorted in descending order of the 'Proportion' column. The first row (row 3) has a proportion of 1.00, which is highlighted with a black border. The second row (row 4) also has a proportion of 1.00. The third row (row 5) has a proportion of 1.00. The fourth row (row 6) has a proportion of .97. The fifth row (row 7) has a proportion of .87. The sixth row (row 8) has a proportion of .86. The seventh row (row 9) has a proportion of .46. The eighth row (row 10) has a proportion of .45. The ninth row (row 11) has a proportion of .33. The tenth row (row 12) has a proportion of .27. The eleventh row (row 13) has a proportion of .22. The twelfth row (row 14) has a proportion of .16. The thirteenth row (row 15) has a proportion of .08. The fourteenth row (row 16) is empty.

	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

This tutorial provides a very brief overview of PDI-CE about data management capabilities. It is possible to define a process with various tasks on the database (loading, transforming and cleaning) without having to write a single line of code for it.