# 1 Topic

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

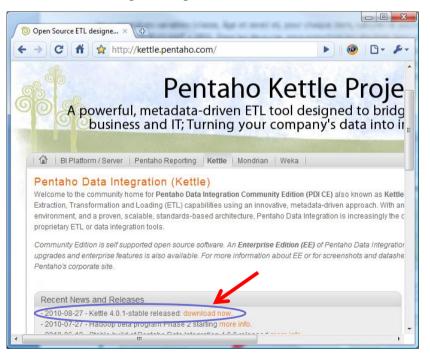
The Pentaho BI Suite is an open source Business Intelligence suite with integrated reporting, dashboard, data mining, workflow and ETL capabilities (<u>http://en.wikipedia.org/wiki/Pentaho</u>)<sup>1</sup>. In this tutorial, we talk about the Pentaho BI Suite Community Edition (CE) which is freely downloadable. 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 <u>TITANIC</u> dataset 32 times in this tutorial (<u>titanic32x.csv.zip</u>). 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> <u>http://www.pentaho.com/</u>
- <sup>2</sup> <u>http://community.pentaho.com/</u>
- <sup>3</sup> <u>http://kettle.pentaho.com/</u>

<sup>&</sup>lt;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.

😋 🔵 🗢 🔰 « pdi-ce-4.0.1-sta	ıble 🕨 data-integration 🔀	<b>▼ \$</b> 9	Rechercher	م
Fichier Edition Affichage O	utils ?			
🌗 Organiser 👻 🏢 Affichages	👻 🛒 Diaporama  😢 Grave		_	0
Nom	Date de la prise de Mots-clé	s Taille	Notation	
🚳 Kitchen.bat		3 Ko	****	
kitchen.sh		2 Ko	****	
🚳 Pan.bat		3 Ko	****	
pan.sh		2 Ko	****	
README_INFOBRIGHT.txt		1 Ko	****	
README_LINUX.txt		1 Ko	***	
README_OSX.txt		2 Ko	***	
README_UNIX_AS400.txt		1 Ko	****	·
🚳 run_kettle_cluster_example.bat		1 Ko	****	
runSamples.sh		1 Ko	****	
🚳 Spoon.bat		3 Ko	****	
S spoon.ico		81 Ko	****	
🕹 spoon.png		3 Ko	****	
spoon.sh		4 Ko	***	

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

🚫 Pentaho Data Integration - Bier	ivenue	
<u>Fichier</u> Edition <u>V</u> ue Action	Tools Aide	
	Astuce du jour	Perspective: 😥 Data Integration
Palette de création »1	B Afficher l'onglet de b	
Fonctions 🛛 🖻 🗂	Afficher information	pl Pentaho Data Integration
	Afficher informations	5 pl
	À propos	Pentaho Data Integration - Spoon version Community Edition - 4.0.1
	-11-	(c) 2001-2009 Pentaho Corporation www.pentaho.org
		Build version : 4.0.1-stable Build date : 2010-08-26 14.18.03
		ОК
	Démarrez Maintenant	Exemples Quoi de neuf?
	Documen tati	on Communauté OBlogs Opentation Desenverse l'instructions et diagence

# 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.

Nouveau	+	Transformation	CTRL-N	🕴 🗛 😥 🕅 🗛 🖓 Perspective: 😥 Data Integrat	tion
Ouvrir	CTRL-O	Tâche	CTRL-ALT-N		
Ouvrir un fichier depuis	une URL	Connexion BDD		gration/docs/Francais/welcome	/kettle
Open Recent	• T	1 /			
Fermer	🚫 Pentaho Data Integratio	on - Transformation 1	1	V	
Close All	Fichier Edition Vue	Action Tools Aide	1		
Enregistrer	1 6 1 8				Perspective: 😥 Data Integration
Enregistrer sous	🔯 Navigateur 🥖 Palett	e de création	🕞 Bienven	ue 🕃 Transformation 1	8
Enregistrer sous (VFS).				📐 🔆 🖻 😡 🗞 🕵	100% ▼
Imprimer	Fonctions				
	Transform		<u>^</u>		
Importer depuis un fic	Extraction				
Export	💫 Chargement conte				
Quitter	Extraction de donn		=		
Quitter	Extraction depuis o			_	
	Extraction depuis I				
	L Extraction depuis f				
	L Extraction depuis f				
	Extraction depuis f				
	V Extraction depuis f				
	Extraction repuis f				
	Extraction depuis f				
		ichier longueur fixe			
	Extraction depuis f				
	🛛 🎇 Extraction depuis F				
	🔹 🔯 Extraction depuis f				

#### 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.

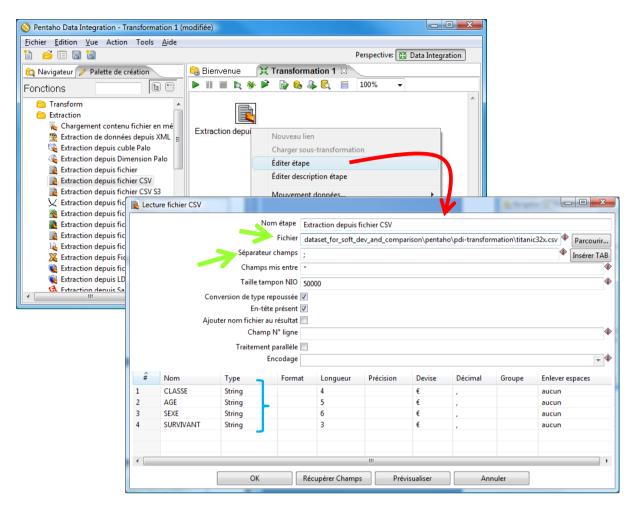
🕙 t	titanic_count.xls	[Mode de con	npatibilité]		-		x
	A	B	С	D	E	F	-
1	CLASSE	AGE	SEXE	SURVIVANT	COUNT_SURVIVANT		
2	1ST	ADULT	FEMALE	NO	128.00		_
3	1ST	ADULT	FEMALE	YES	4 480.00		
4	1ST	ADULT	MALE	NO	3 776.00		
5	1ST	ADULT	MALE	YES	1 824.00		
6	1ST	CHILD	FEMALE	YES	32.00		
H I	Sheet	1 🕲 🗌		I 4 📃		•	١.

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>&</sup>lt;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.

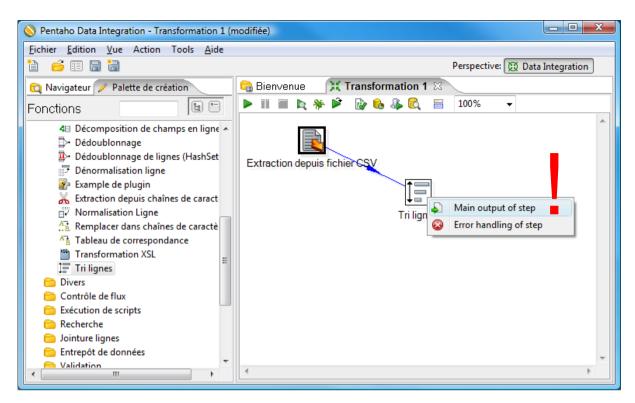


To check of the data reading, we click on the "**Récupérer Champs**" button. The tool determines automatically the column data type. It uses the 100 first rows for that. We can modify this setting. In our dataset, the categorical variables are defined as STRING.

### 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.

📮 Tri lig	Ines	the contribution from	and the second in	
	Nom étap	e Tri lignes		
	Répertoire de t	ri %%java.io.tm	pdir%%	Parcourir
	Préfixe fichiers TM	IP out		
Nł	or lignes en mémoi	re 1000000		\$
Cor	nserver min mémoi	re		\$
	Compresser fichier	rs 🕸 📃		\$
	Transmettre valeu	rs 📃		
Champs	:			
÷	Nom champ	Ascendant	Respecter la casse	
1	CLASSE	0		
2	AGE	0		
3	SEXE	0		
4	SURVIVANT	0		
	ОК	Annu	ler Récupére	er Champs

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).

	2 Agrégation	
	Nom étape Agrégation valeurs	
Nentaho Data Integration - Transformation 1 (modifiée)	Inclure toutes les lignes 🗐	
Eichier Edition Vue Action Tools Aide	Répertoire des fichiers temporaires %%java.io.tmpdir%%	Parcourir
🗎 🧉 🗐 🔚 Perspective: 👿 Day	Préfixe fichier TMP grp	
🔯 Navigateur 🥖 Palette de création 💦 🕞 Bienvenue 🛛 💥 Transformation 1 🕸	Ajouter numéro de line, initialiser à	
Fonctions	Nom champ contenant numéro	
Extraction	Toujours retourner une ligne	
C Alimentation	Les champs de groupement:	
Divers	Champ groupe	Récupérer champs
Contrôle de flux	1 CLASSE	
💼 Exécution de scripts		
Contraction E E E E E E E E E E E E E E E E E E E	2 AGE 3 SEXE (1)	
Entrepôt de données	4 SURVIVANT	
Calidation	Agrégation :	
C Statistiques	# Nouveau champ Champ agrégé Type	Récupérer champs
Page Agrégation valeurs	1 COUNT SURV SURVIVANT Nombre de valeurs	*
Echantillonnage aléatoire		
■* Echantillonnage fixe		
2 Memory Group by		
🖟 Statistiques simples 🗸 🤟		
	(2)	
	<	- F-
	OK Annuler	

#### 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).

Extraction depuis fich	Tri lignes	grégation vale	Alimentation	fichier Xcel
Alimentation fichier Excel		Ľ		
Nom étape Fichier Contenu Champs	Alimentation fich	ier Excel		
Nom fichier Ne pas créer le fichier au Extension Inclure n°étape dans nom de fichier	xls	odi-transformation\t	iitanic_count 🏶 Pa	arcourir
Inclure date dans nom de fichier	Alimentatic	on fichier Excel		
Inclure heure dans nom de fichier	Annendade	AT THEMET EXCEL		
Spécifier format date Format date		Nom étape Ali	mentation fichier l	Excel
i offiac date	Fichier Conte	enu Champs		
	<b>a a a b b b b b b b b b b</b>	m	Туре	Format
Ajouter fichier au résultat	1 CL	ASSE	String	
	2 AG	E	String	
	3 SE		String	
		RVIVANT	String	
	5 CC	OUNT_SURV	Integer 🤳	
		Récupérer Cha	amps Large OK Annu	ur Minimale

#### 4.1.4 Launching the calculations

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

X Exécuter une transformation		
Locale, à distance ou en grappe	e	
Exécution locale	Exécution à distance	Exécution en grappe
	Hôte distant	
	Envoyer toute la trans	formation vers le serveur l'exécution
	-	Démarrer l'exécution
		Afficher les transformations
Détails		
	C Act	tiver le mode Contrôle renforcé
	📝 Vid	er la trace avant l'exécution
Niveau de la trace	Nivea	au de base 👻
Date de ré exécution (aaaa/MN	I/jj HH:mm:ss)	
Paramètres	Variab	les
# Paramètre Valeu	ir Valeur par défa #	Variable
	· 1	Internal.Job.Filename.Directory
Arguments	2	Internal.Job.Filename.Name
# Argument Valeur	3	Internal.Job.Name
1		4
_	Démarrer A	Innuler

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

🚫 Pentaho Data Integration - test		
<u>Fichier Edition Vue</u> Action Tools	Aide	
1 🖻 🗉 🖬 🔚	Perspective: 👿 Data	Integration
Palette de création <sup>3</sup> 1	🕞 Bienvenue 🛛 💥 test 🛛	
Fonctions 🛛 🕫 🖱	🕨 🏢 🖿 📐 🎋 🖻 🔯 🗞 🥾 🕄 🐁 100% 🗸	
Transform Extraction Alimentation	Tri lignes Alimentation fichier Excel	E
Alimentation cube Palo ■	Extraction depuis fichier CSV Agrégation valeurs	-
Alimentation dimension Pa Alimentation fichier	4	- F
Alimentation fichier Excel	Résultats exécution	
Alimentation fichier kettle	A Historique	
Alimentation fichier XML	<b>Z</b> <sub>22</sub>	
Export instructions SQL ver	🗘 Nom étape N°Copie Lignes lues L	Lignes écrites
S Génération flux RSS	1 Extraction depuis fichier CSV 0 0	70432
Insertion / Mise à jour table	2 Tri lignes 0 70432	70432
Insertion dans Salesforce Insertion dans table	3 Agrégation valeurs 0 70432	24
A Insertion/Mise à jour Salesf	4 Alimentation fichier Excel 0 24	24
A Mise à jour de Salesforce ▼	< <u> </u>	4

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

	А	В	С	D	E	
1	CLASSE	AGE	SEXE	SURVIVANT	COUNT SURV	
2	1ST	ADULT	FEMALE	NO	128.00	ר
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	П
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	J

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.

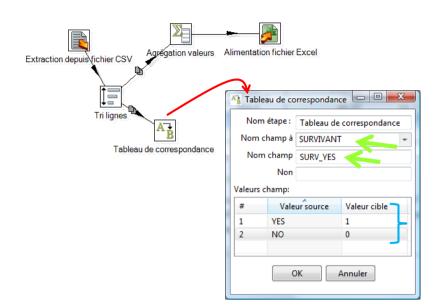
	titanic_freq - va	lidated.xls [Mod	e de compatibili	té]	_	= x
	A	B	С	D	E	F 🚍
1	CLASSE	AGE	SEXE	COUNT_SURVIVANT	FREQ_SURV_YES	
2	1ST	ADULT	FEMALE	4 608.00	97.22%	
3	1ST	ADULT	MALE	5 600.00	32.57%	
4	1ST	CHILD	FEMALE	32.00	100.00%	
5	1ST	CHILD	MALE	160.00	100.00%	<b>•</b>
H	♦ ► ► Sheet	1 🖓		[ ◀		► <b>I</b>

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, o otherwise. Thus, when we compute the mean of this new column, we obtain the proportion of SURVIVANT=YES.

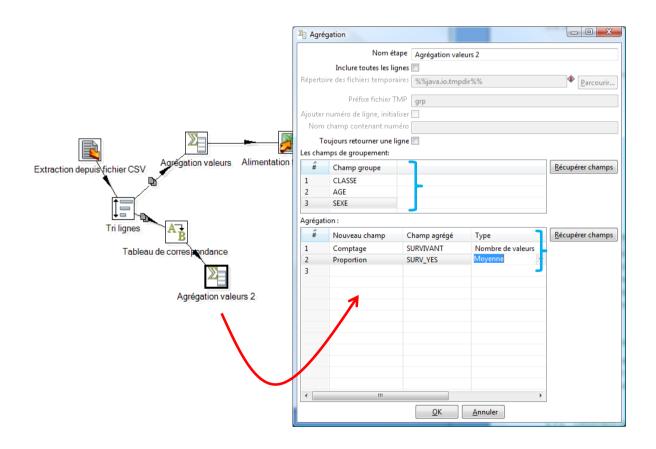
To do this, we use the **"Tableau de correspondance**" component. We connect the sorted array (this avoids to redo twice the sorting operation afterwards). We set the following settings.



#### 4.2.2 Aggregating the rows

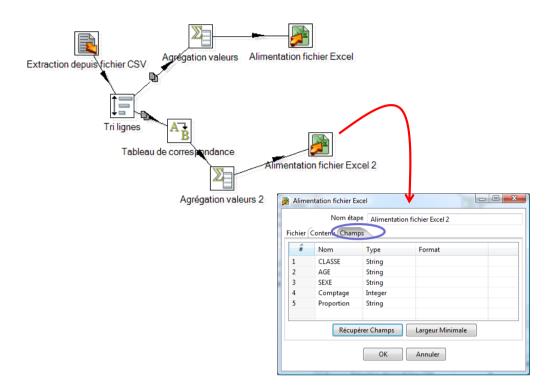
We specify the analysis with the "**Agrégation valeurs**" component. Two calculations are needed for each occurrence of the itemsets incorporating CLASSE, AGE and SEXE: (1) counting the corresponding instances, (2) computing the frequencies of SURVIVANT = YES.

We set the following parameters into the dialog settings.



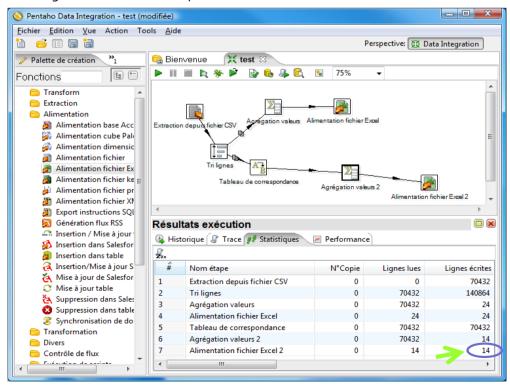
### 4.2.3 Exporting the results

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  $\triangleright$  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.

	A	В	С	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.