



# PRESENTATION DE PENTAHO DATA INTEGRATION (PDI)

*Août 2006 – Version: 1.0*  
*Auteur: Samatar HASSAN*

**PDI 2.3**  
**<http://www.pentaho.org/>**

## PLAN

<b>I Présentation .....</b>	<b>3</b>
<b>1.1 Définition .....</b>	<b>3</b>
<b>1.2 La petite histoire.....</b>	<b>3</b>
<b>1.3 Les composants PDI.....</b>	<b>3</b>
<b>1. SPOON.....</b>	<b>4</b>
<b>2. PAN.....</b>	<b>5</b>
<b>3. CHEF.....</b>	<b>6</b>
<b>4. KITCHEN.....</b>	<b>7</b>
<b>II Installation.....</b>	<b>8</b>
<b>2.1 Prérequis .....</b>	<b>8</b>
<b>2.2 Documentation.....</b>	<b>8</b>
<b>2.3 Participez à l'aventure .....</b>	<b>8</b>

## I Présentation

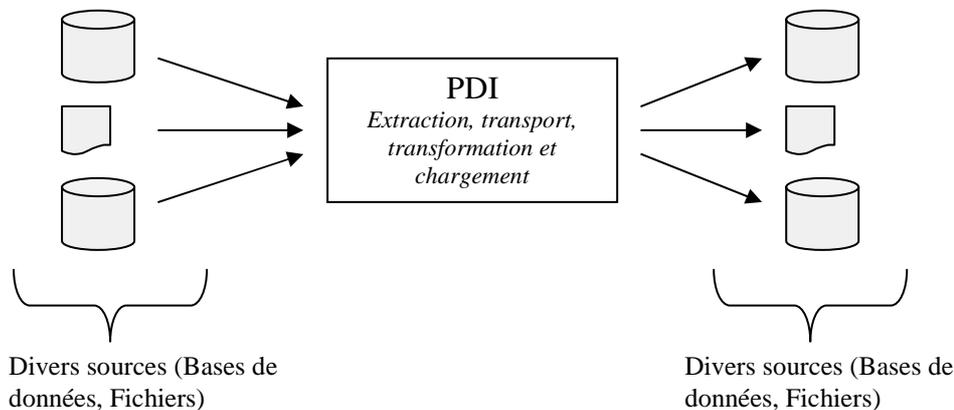
---

Dans ce document, nous allons présenter très brièvement l'outil ETL Open Source Pentaho Data Extraction (PDI).

### 1.1 Définition

Pentaho Data Integration (anciennement K.E.T.T.L.E – Kettle ETL Environment) est un E.T.T.L, c'est-à-dire qu'il permet :

- L'**E**xtraction des données depuis divers source (fichiers, bases de données)
- Le **T**ransport des données d'une unité de stockage à une autre
- La **T**ransformation des données
- Le chargement (**L**oading en anglais) des données dans un entrepôt



Ce produit Open Source fournit une interface graphique pour la manipulation des données et cela contrairement à la plupart des autres produits non commerciaux.

### 1.2 La petite histoire

KETTLE a été développé il y a 5 ans par Matt CASTERS, un consultant en Business Intelligence (BI) indépendant, dans un premier temps pour ses propres besoins.

Le projet a été rendu open Source l'année dernière et PENTAHO l'a acquis au début de l'année 2006. C'est ainsi que KETTLE est devenu Pentaho Data Integration (PDI).

Matt conserve le leadership sur le projet en tant que « Chief Data Integration » chez PENTAHO.

Intéressons nous maintenant au produit lui-même.

### 1.3 Les composants PDI

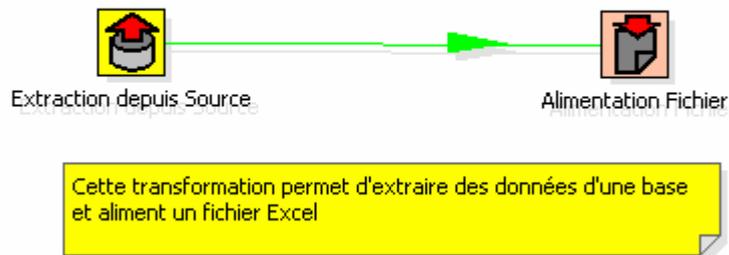
Comme nous l'avons vu plus haut, PDI est un environnement qui permet d'une part de définir des transformations sur les données, de les exécuter et d'autre part de les sauvegarder dans des fichiers ou dans un référentiel base de données.

De plus, PDI permet de connecter à un grand nombre de bases de données commerciaux ou non.

Ainsi plusieurs outils composent cet environnement :

1. **SPOON** est l'outil qui permet grâce à son interface graphique de créer des transformations, les exécuter et les sauvegarder.  
Les composants permettant la manipulation des données sont nommés « **étapes** » (steps en anglais).

Par exemple il existe une étape permettant d'extraire des données de diverses bases de données, une autre aidant à l'extraction depuis des fichiers.  
SPOON comprend un grand nombre d'étapes.



L'exemple ci-dessus a été créé grâce à SPOON. La ligne verte entre les étapes est un **lien** (Hop en anglais). C'est lui qui indique vers quelle étape est dirigé le flux (matérialisé par le sens de la flèche).  
Notons à ce stade que SPOON manipule des enregistrements (ou lignes) sous la forme suivante :

Colonne 1	Colonne 2	.....	Colonne n
VC10	VC20		VCN0
VC11	VC21		VCN1

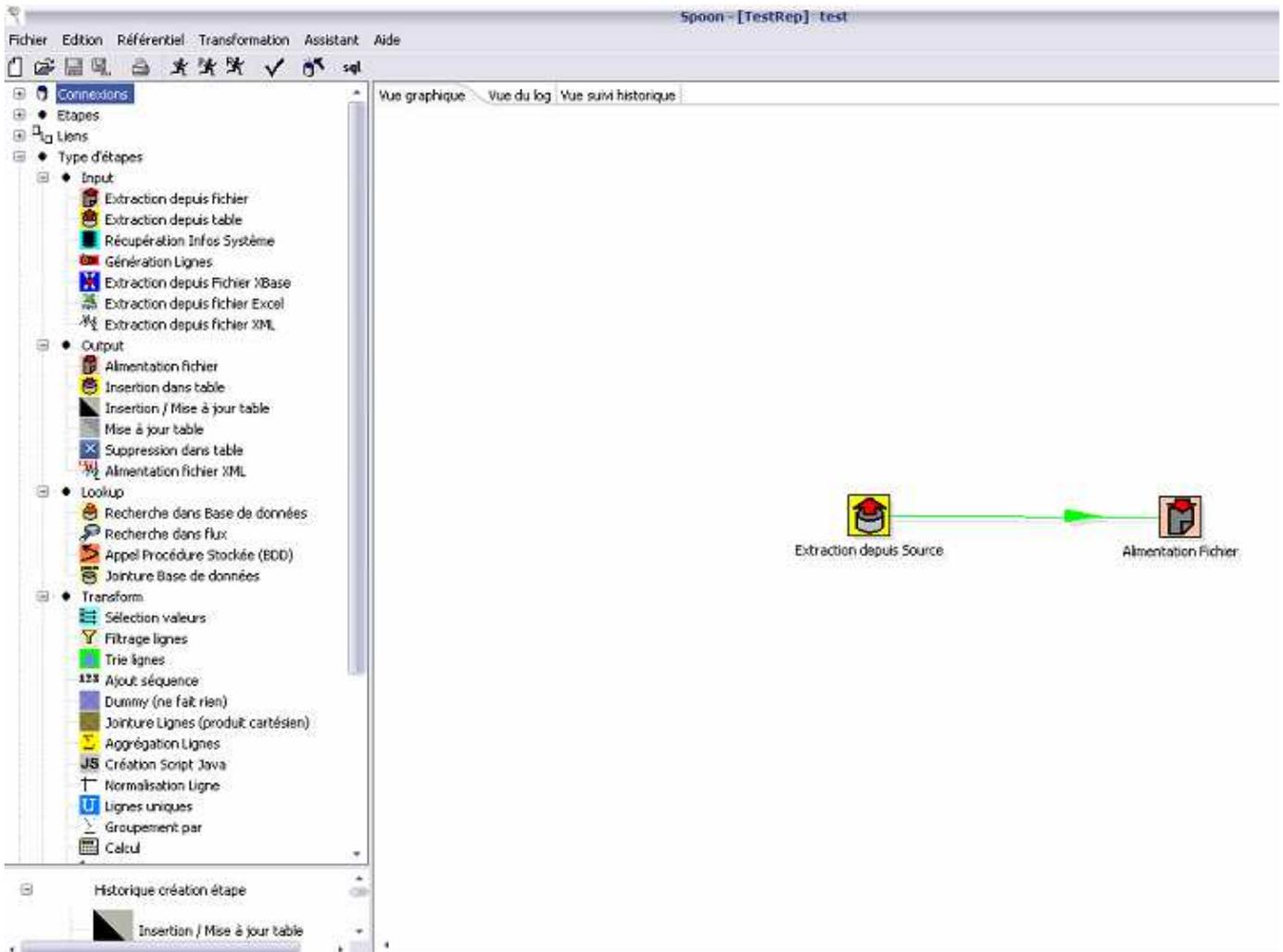
...

Grâce à SPOON, vous pourrez donc créer vos transformations, les tester et les sauvegarder soit dans un fichier, soit dans un référentiel d'une base de données que vous aurez préalablement créée.

L'écran suivant donne un aperçu de l'interface de SPOON. Les différentes étapes sont visibles dans la partie gauche de l'interface.

Le schéma de la transformation est dans la partie droite de l'interface.

Les étapes sont simplement déposées sur la partie droite (drag & drop) à partir de la partie gauche.



Mais vous voulez certainement pouvoir automatiser l'exécution de votre transformation à des horaires de votre choix.

C'est à ce niveau qu'intervient PAN.

2. **PAN** est un outil, très simple d'utilisation, permet d'exécuter une transformation en ligne de commande. Ensuite on pourra planifier l'exécution grâce par exemple au planificateur de Microsoft Windows ou un Cron dans l'environnement Unix.

Lorsque vous devez alimenter un entrepôt de données, vous avez à exécuter plusieurs transformations (extraction des dimensions, alimentation des faits,...). Ces transformations ne sont pas indépendantes les unes des autres. En effet, l'alimentation des tables de faits ne doit être réalisée que si les données de dimension ont été insérées avec succès dans l'entrepôt, or SPOON n'a pas pour vocation de gérer ni la séquentialité des transformations, ni le fait qu'une transformation s'effectue avec succès.

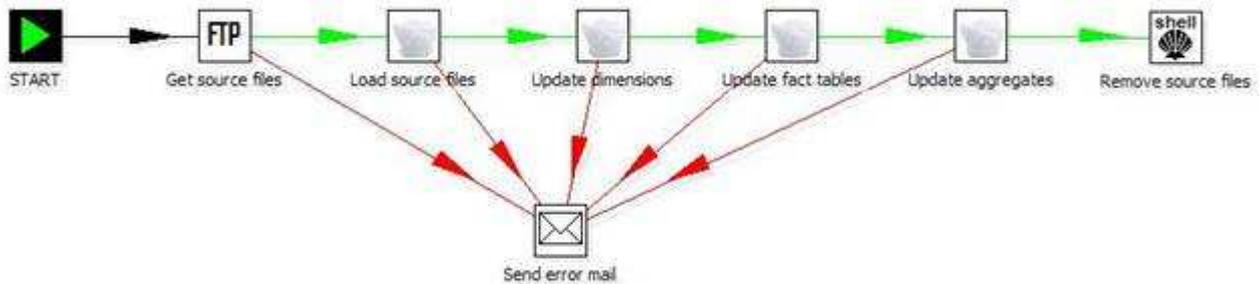
Nous introduiront un autre outil : **CHEF**.

**3. CHEF introduit une autre notion : La tâche** (ou Task en anglais).

Une tâche est une organisation qui permet d'automatiser des tâches complexes de transformations.

En effet, l'exécution de chaque entrée ne démarre que si l'entrée précédente a été terminée. De plus, on peut être le résultat de chaque entrée. A-t-elle été exécutée avec succès ?

Une entrée peut être une transformation ou des transformations spéciales comme la récupération de fichiers par FTP ou l'exécution de fichier shell...



Commentons l'exemple ci-dessus.

Listons toutes les entrées de la tâche :

- L'entrée « Start » indique le démarrage de la tâche (on n'en trouve qu'une seule par tâche).
- L'entrée « Get source files » permet de récupérer des fichiers depuis un serveur FTP. Les fichiers ainsi obtenus sont stockés dans un répertoire.
- Les entrées « Load source files », « Update dimensions », « Update fact tables », « Update aggregates » exécutent des tâches (sous-tâches)
- L'entrée « Remove source files » permet de supprimer les fichiers récupérés.

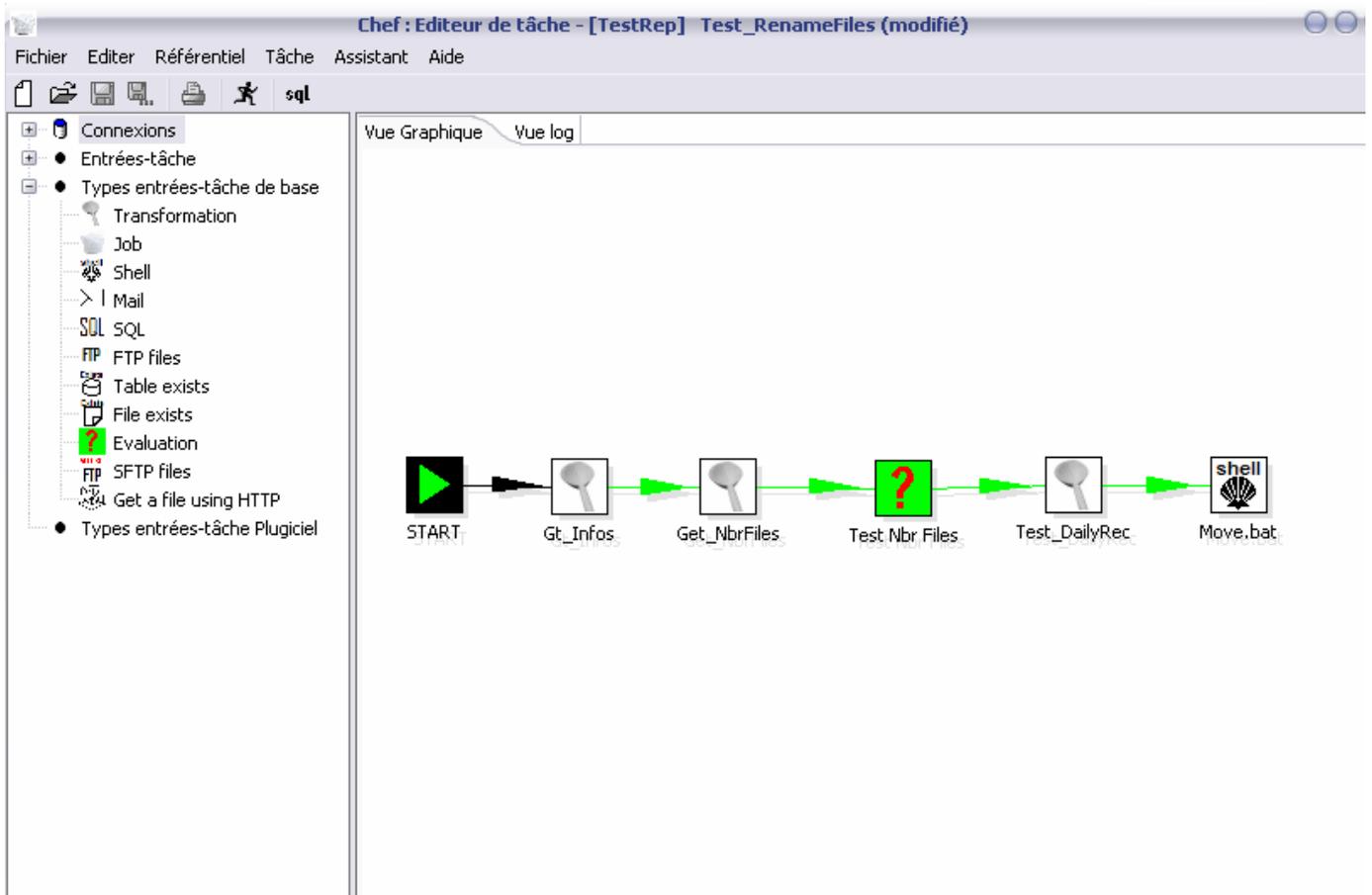
Observez maintenant les flèches **vertes** entre les entrées. Cela indique que l'étape suivante ne va être exécutée que si l'entrée précédente s'est bien déroulé (elle n'a pas généré d'erreur).

La dernière entrée en « Send error mail ». Un courriel est envoyé si une entrée est en échec (flèche **rouge**).

**CHEF** fournit une interface graphique permettant la création, l'exécution et la sauvegarde des tâches.

Ainsi chef vous permettra par exemple de surveiller l'exécution de vos transformations. Vous pouvez décider d'envoyer un courriel à une personne si la transformation a généré des erreurs.

L'image suivante montre un aperçu de l'interface de CHEF.



A l'instar de SPOON, un outil en ligne de commande est disponible pour CHEF.

#### 4. KITCHEN

KITCHEN permet d'exécuter une tâche en ligne de commande.

## II Installation

---

### 2.1 Prérequis

Pour fonctionner, PDI a besoin de l'environnement d'exécution JAVA.  
Vous devez donc si ce n'est pas encore le cas, installer la machine virtuelle Java 1.4 ou au dessus.  
Cet outil est téléchargeable gratuitement sur le site <http://www.javasoft.com>.

Une fois cette étape effectuée avec succès, il suffit de se procurer la dernière version de PDI : 2.3 sur le site De PENTAHO  
<http://prdownloads.sourceforge.net/pentaho/Kettle-2.3.0.zip?download>

Les dernières mises à jour sont disponibles sur le site : [http://www.javaforge.com/proj/doc.do?proj\\_id=318](http://www.javaforge.com/proj/doc.do?proj_id=318)

Une fois le précieux fichier zip récupéré, il suffit de le dézipper dans le répertoire de votre choix.

Selon votre environnement (Windows ou Unix) lancer le fichier SPOON.bat (windows) ou SPOON.sh (Unix) pour démarrer SPOON et CHEF.bat (ou CHEF.sh) pour démarrer CHEF.

### 2.2 Documentation

La documentation est également fournie (dans le répertoire docs) certes pour l'instant en anglais : La traduction en français suivra.

N'hésitez pas à la consulter car elle est très bien faite. Si toutefois vous recherchez de l'aide, rendez-vous sur le forum :

[http://www.javaforge.com/proj/forum/browseForum.do?forum\\_id=1274](http://www.javaforge.com/proj/forum/browseForum.do?forum_id=1274)

### 2.3 Participez à l'aventure

PDI est en perpétuelle amélioration. Vous avez la possibilité de signaler des bugs éventuels à l'adresse suivante :

[http://www.javaforge.com/proj/tracker/browseTracker.do?tracker\\_id=1273](http://www.javaforge.com/proj/tracker/browseTracker.do?tracker_id=1273)

Vous avez une idée, n'hésitez pas à la partager avec les utilisateurs de l'outil et postez une amélioration à l'adresse suivante :

[http://www.javaforge.com/proj/tracker/browseTracker.do?tracker\\_id=1274](http://www.javaforge.com/proj/tracker/browseTracker.do?tracker_id=1274)