# ETL architecture – Choix de l’ETL

Sommaire

[ETL architecture – Choix de l’ETL 1](#_Toc286630120)

[Introduction 2](#_Toc286630121)

[Comparatif des logiciels 3](#_Toc286630122)

[Les critères de sélections 3](#_Toc286630123)

[Le prix 3](#_Toc286630124)

[La communauté d’utilisateurs 4](#_Toc286630125)

[La renommée 4](#_Toc286630126)

[L’ergonomie 4](#_Toc286630127)

[Les fonctionnalités 5](#_Toc286630128)

[Choix final et conclusion 5](#_Toc286630129)

LE QUANG Sylvain

BERNARD Valentin

## Introduction

Pour transposer les fichiers Excels mensuels et annuels par le client dans la base de données, nous devons avoir recours à l’alimentation et plus particulièrement des outils de type Extract Transform Load (appelé aussi ETL). Ces outils nous permettront donc d’importer les données dans le fichier Excel dans une base de données en transformant les données dans un le fichier Excel en un flux qui sera compatible avec la base de données destinée. Ce procédé est ce qu’on appelle aussi le mapping de données.

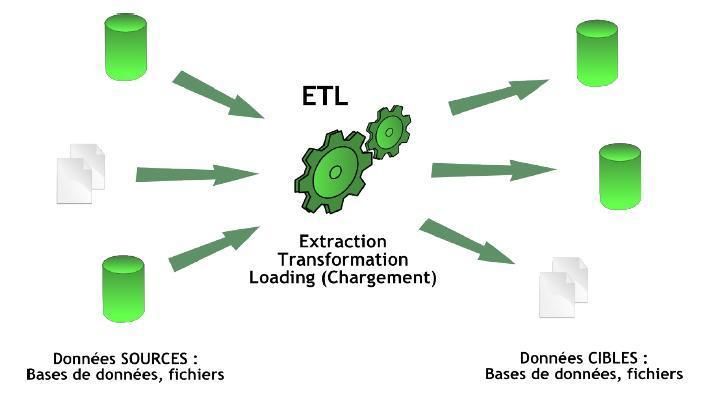


Figure 1 Schéma de l'ETL

Pour choisir l’outil ETL qu’on utilisera tout au long du projet, nous avons comparé plusieurs logiciels de type ETL afin de choisir celui qui conviendrait le plus à notre type de projet.

## Comparatif des logiciels

### Les critères de sélections

Nous avons tenté de classer des ETL selon des critères qui seraient propres à notre projet, nous avons fait un brainstorming concernant les critères importants selon nous et il en est ressorti les critères suivants :

* Prix
* La communauté d’utilisateurs
* La renommée
* Les fonctionnalités
* Le choix des autres groupes
* L’ergonomie

Ces critères ont bien sûr un poids différent, par exemple le prix est un critère plus important que les autres, car si on veut plus ou moins de budget, l’accord du client est primordial.

### Le prix

Tout d'abord, nous avons exclu les logiciels payants, car même s’ils sont souvent complets, ils s’adressent tout d’abord à des professionnels de l’ETL. Étant donné que nous souhaitons découvrir et nous initier à l’ETL et que nous n’avons pas de budget, nous nous sommes plutôt tournés vers des solutions open source qui proposent des versions gratuites.

En cherchant des informations on a pu voir quelques acteurs sur le marché de l’ETL open source, nous avons :

* Talend Open Studio 4.1.1
* JasperETL 3.2.3
* SpagoBI 2.3
* Pentaho Data Integration 4.1 (appelé aussi Kettle)

Ces solutions ont été trouvées à partir de recherche sur Internet, notamment un livre blanc d’Atol Consulting [[1]](#footnote-1)qui compare les fonctionnalités et performances de Talend Open Studio et Pentaho Data Integration. Nos recherches ont été aussi complétées en côtoyant des forums Internet comme hardware.fr qui possèdent une forte communauté d’informaticiens qui ont des conseils et des avis concernant l’ETL à utiliser pour notre projet.

Il faut aussi noter que nous avions une solution SAS à notre disposition grâce à la fac qui dispose des licences, mais étant donné l’instabilité des serveurs à l’ISTIL ainsi que de la version SAS que l’ISTIL avait, nous nous sommes tournés vers d’autres solutions.

### La communauté d’utilisateurs

Une grande communauté d’utilisateurs est importante, car tout d’abord, elle assure la pérennité du logiciel, car un logiciel sans communauté qui l’utilise montrera qu’il y a un problème avec ce logiciel.

De plus, une communauté permet d’avoir une meilleure documentation du logiciel, que ce soit au niveau tutoriel, mais aussi dans la résolution de problème. Cela permet une meilleure prise en main si on est accompagné tout à long de notre initiation à l’ETL.

De plus, une communauté d’utilisateurs importante montre que le logiciel ETL répond à des besoins et que beaucoup de gens croient en l’ETL.

En étudiant le livre blanc d’Atol Consulting et les différents sites de constructeur, Talend et Pentaho ont les meilleures communautés d’utilisateur avec un petit plus pour Talend, car grâce à son Talend Forge ou ce sont les utilisateurs qui créent et valident eux-mêmes de la documentation et tutoriel, on retrouve beaucoup de tutoriels en multi langue, ce qui est appréciable si on n’est pas habitué a l’anglais.

On peut aussi noter qu’il existe un support professionnel pour Talend et Pentaho, mais celui-ci est disponible seulement dans la version payante, cependant il est bon de savoir qu’il existe aussi un support professionnel, cela est rassurant au cas ou la communauté n’a pas réponse a un problème qu’on pourrait rencontrer.

### La renommée

La renommée est aussi importante, car cela montre le rayonnement et la reconnaissance du logiciel dans le monde. Talend a été reconnu comme l’ETL open source le plus efficace en consultant divers forums.

De plus, il faut savoir que JasperETL est en fait Jasper qui a intégré la solution Talend dans leur suite logiciel, de ce fait, Talend inspire confiance, car même d’autres éditeurs intègre la solution Talend et nous rassure sur les performances et fonctionnalités du logiciel.

### L’ergonomie

Pentaho et Talend se basent tous les deux sur le langage JAVA et il faut donc des notions pour débuter et comprendre le logiciel, car il y a beaucoup de références au java lors de l’utilisation, cependant Pentaho a une interface propriétaire alors que Talend utilise la même interface qu’Eclipse, de ce fait nous serons avantagés en utilisant Talend, car nous utilisons régulièrement Eclipse lors de nos projets scolaires.

### Les fonctionnalités

Pour les fonctionnalités, nous avons besoin des fonctionnalités de base, c'est-à-dire que l’ETL puisse ouvrir des fichiers Excel, soit compatible avec une base Oracle v11 et soit capable de générer le mapping des données.

Vu que c’est le role d’un ETL, tous nos ETL ont ces fonctionnalités, cependant il faut noter que Talend est un logiciel de type « générateur de code » alors que Pentaho est du type « moteur de code ».

En d’autres termes Talend génère du code Java et de ce fait on peut exécuter tous nos jobs de mapping de donnée sans avoir Talend d’installer sur le serveur, ce qui n’est pas le cas de Pentaho.

Cela facilite donc l’exécution des jobs sur le serveur puisqu’on n’a pas besoin de Talend pour lancer nos jobs.

## Choix final et conclusion

Au final, il en ressort que Pentaho et Talend sont de bon ETL. Nous avons choisi Talend car il répond mieux aux critères que Pentaho et il nous semble bien plus accessible pour une initialisation à l’ETL. De plus, on constate que Talend a une forte volonté de s’implanter en France en organisant régulièrement des journées de formation dans toutes les grandes villes (il y en a eu à Lyon fin 2010).

1. Le livre blanc est disponible à l’adresse suivante : <http://blog.atolcd.com/?tag=talend> [↑](#footnote-ref-1)