Cours: ENTREPOT DES DONNEES

Responsable du Cours : S. DAMMAK

Auditoire : 2LGSI
A-U : 2020-2021

Talend open studio For data integration

I. Talend

Talend est une société française créée en 2005, développant une suite de nombreux logiciels Open Source, connue sous le nom de **Talend Open Studio**.

« Talend Open Studio » **TOS** est un ensemble de produits open source pour le développement, test, déploiement et administration des projets d'intégration de données et d'applications. TOS fournit une plateforme unifiée qui rend la gestion et l'intégration des données et applications plus facile, en fournissant un environnement unifié pour la gestion de tout leur cycle de vie.

Il existe plusieurs solutions offertes par Talend :

- **Big Data** : Environnement qui facilite la gestion des données volumineuses.
- **Data Integration** : Ensemble d'outils pour l'intégration de données pour accéder, transformer et intégrer les données à partir d'un système en temps réel pour remplir les besoins d'intégration des données.
- **Data Quality** : Permet d'assurer le profiling et monitoring des données pour identifier des anomalies et assurer la qualité des données.
- **ESB** : Permet la création, la connexion, la médiation et la gestion des services et leurs interactions.

➢ *Avantages :*

- Le logiciel est **gratuit avec un code source mis à disposition** (avec de nombreuses fonctionnalités supplémentaires payantes) ;
- Le résultat graphique permet une meilleure visibilité des jobs
- De ce fait la **productivité des développeurs** est accrue : le codage est plus rapide qu'en SQL malgré le temps de prise en main ;

- TOS for Data Integration met actuellement à disposition plus de 600 composants, ce qui en fait un **outil puissant et varié** ;
- Il est possible de **créer ses propres composants visuels** et de les diffuser avec la communauté ; l'enrichissement de la bibliothèque de composants se fait donc naturellement et par tous ;
- La construction des requêtes facilitée dans les bases de données en détectant le schéma et les relations entre tables ;
- Il existe également des **fonctionnalités complémentaires** : il est par exemple possible de réaliser depuis le logiciel des schémas expliquant les flux, de stocker des documents (PDF, JPG.....) au sein de l'espace de travail.

➢ ***Inconvénients :***

- Le logiciel présente quelques défauts d'ergonomie. Il est par exemple difficile de naviguer dans la palette en cherchant un composant précis sans utiliser le moteur de recherche ;
- Le mode graphique avec uniquement des clics a ses limites : comment communiquer avec un collègue/sur un forum d'entraide, et expliquer clairement la démarche que nous avons suivie ?
- Le logiciel très lourd, tout comme l'environnement Eclipse qu'il utilise.

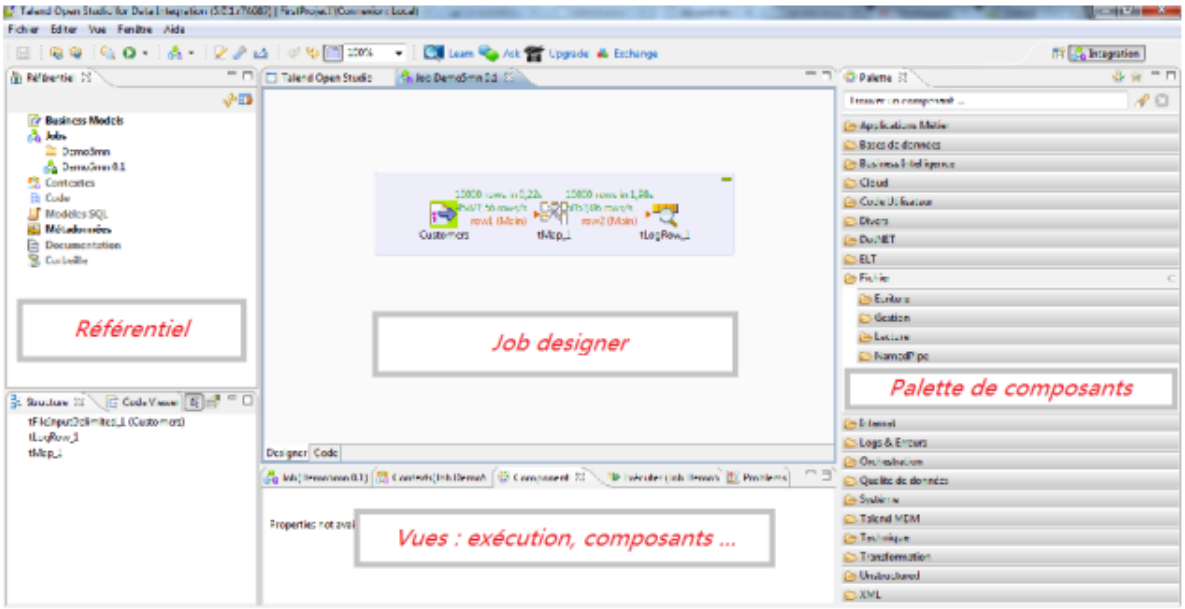
II. Intégration des données

L'intégration des données est activité consistant à regrouper les données issues de différentes sources, les unifier (i.e. modifier pour les rendre cohérentes entre elles) pour donner à l'utilisateur une seule et même vue des données disponibles.

Talend Open Studio for Data Integration est solution **ETL** (Extract, Transform and Load) Open Source permettant de répondre avec efficacité à un très large éventail de besoins : traiter les données volumineuses à partir de données en entrée appelées source vers des destinations SGBD ou fichiers (csv, txt, xml …) appelées cibles.

Pour les besoins de notre TP, nous utilisons « Talend Data Integration » pour la transformation des données et leur intégration. Il est possible de télécharger toutes les solutions de TOS sur http://fr.talend.com/products/talend-open-studio.

III. Interface et fonctionnalités :



Au centre se trouve le "job designer", c'est ici que vous pouvez agencer graphiquement vos différents composants pour construire vos jobs.

A gauche, la partie « référentiel » regroupe tous les éléments que vous avez importé ou créé par l'utilisateur: c'est à cet endroit que vous retrouverez vos fichiers délimités liés à vos imports, et vos routines pour les fonctions Java que vous aurez écrites vous-mêmes par exemple.

A droite se trouve la palette de composants, organisés en répertoire.

Enfin c'est en bas au centre que se trouvent les différentes vues : vue d'exécution, variables de contexte, configuration de certains composants ...

Job

Un job est un ensemble de tâches, considérées comme une unité. Il regroupe de nombreux composants liés entre eux. Il peut être exécuté en mode batch in interactif, avec de multiples inputs/outputs autorisés au sein d'un même job.

Les jobs peuvent être synchronisés entre eux de diverses façons. De plus les jobs peuvent être hiérarchisés : des jobs principaux peuvent lancer des sous-jobs

Composant

Un composant est un sous-ensemble d'un job offrant une panoplie de fonctions. Il peut être considéré comme une unité de traitement (exécution d'une fonction précise). Il en existe environ 600 sous TOS 5.x.

L'objectif des composants est d'éviter la rédaction de lignes de code simples les plus courantes ; (exemple : mettre en majuscule tout une chaîne de caractères)

Ex : tFileInputDelimited (csv en entrée)-> tMap (composant de transformation) -> tLogRow (affichage sur la console des données en sortie)

Connecteurs

Plus de 450 connecteurs sont disponibles pour se connecter aux principaux SGBD (Oracle, PostgreSQL, MySQL,...) ainsi que pour traiter tous les types de fichiers plats (CSV, Excel, XML), aussi bien en lecture qu'en écriture.

Métadonnées

Les métadonnées sont des données qui définissent les données traitées. Un référentiel permet de les stocker afin de pouvoir les exploiter dans différents jobs.

L'objectif des métadonnées est d'obtenir une application plus robuste, plus facile à maintenir, plus rapide

Ex : On peut sauvegarder le type et le format des données d'entrée d'un fichier CSV afin de pouvoir les exploiter ultérieurement.