**Entrepôts de données**

**TP n°1 : Intégration des données avec Talend**

### Rappel :

Le **Système d’Information (SI) est un élément central** d’une entreprise ou d'une organisation. C’est un ensemble de ressources *matérielles*, *humaines* et *logicielles* permettant aux différents acteurs de véhiculer des informations et de communiquer. Ce système permet la création, la *collecte*, le *stockage*, le *traitement* et la *modification* des informations sous divers formats.

Les quantités de données dans ce système sont énormes. Elles proviennent de différentes bases de productions: département commercial, marketing, ressources humaines, facturation, comptabilité, contrôle de gestion, etc. Les données d’analyse se trouvent ***éparpillées*** dans des systèmes multiples, ***déstructurées*** et ***hétérogènes***.

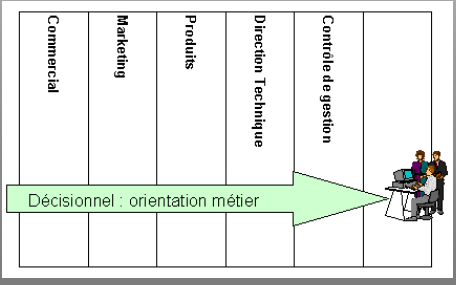


Figure 1. Vue transversale des métiers de l’entreprise: Analyses décisionnelles

Les directeurs opérationnels et les décideurs d'une entreprise ont un besoin crucial sur le déroulement de leurs activités et ils cherchent à répondre à des questions du genre:

* Qui sont mes clients ?
* Comment cibler ma clientèle ?
* Pourquoi le CA baisse ?
* Comment augmenter et améliorer mon CA ?
* Quel est l’évolution d’un tel produit ?

etc***.***

**L'informatique décisionnelle ou Business Intelligence (BI)** offre et représente une opportunité pour les entreprise afin d’optimiser le pilotage de leurs activités. Elle permet une sélection des informations opérationnelles pertinentes pour l'entreprise. Celles-ci sont ensuite normalisées pour alimenter un **entrepôt de données**. La modélisation multidimensionnelle permet de rendre les données d'un entrepôt cohérentes, lisibles, intelligibles et faciles d'accès.

L'informatique décisionnelle doit produire des indicateurs et des rapports à l'attention des analystes. Elle doit également proposer des outils de navigation, d'interrogation et de visualisation de l'entrepôt.

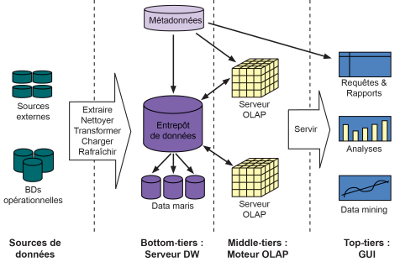


Figure 2. Architecture décisionnelle

Le **processus ETL (Extract Transform Load)** est la phase la plus importante et la plus longue dans le développement d’un projet décisionnel (jusqu’à 70% du temps total). Cette phase permet la *détection des anomalies des données,* de *les transformées* et d’avoir à la fin des *informations pértinentes*, *fiables*, *claires* et *ayant une signification* *logique* ce qui **aide les décideurs** d’une entreprise à prendre la bonne décision dans des délais raisonnables.

ETL (Extract Trnasform Load)



A travers le processus **ETL,** les données sont ***extraites*** à partir des sources hétérogènes (*Extract*), ***consolidées*** (*Transform*) puis ***chargeés*** *après intégration* dans un référentiel unique qui est l’***entrepôt de données.***

1. **C’est quoi Talend Open Studio for Data Integration ?**

* **Talend** est un logiciel open source permettant la gestion des données au sein d’un système d’information (SI), l’échange inter-applicatif de données et la *migration* d’outils métier. Cet outil est régulièrement utilisé pour la mise en place d’***entrepôts de données*** dans le cadre de l’informatique décisionnelle.
* Les sources des données peuvent être une *base de données (BD)*, un *fichier CSV,* un fi*chier TXT*, un *service Web,* etc.

Cet outil intervient lors du *processus d’intégration des donnée*s à une base de données relationnelle qui sera utilisée comme un ***entrepôt de données*** (composé d’une ou plusieurs ***tables de faits*** entourées de plusieurs ***tables de dimension***), destiné aux processus analytiques de l’entreprise : *reporting*, *tableaux de bords*, *fouille de données*, etc.

1. **Objectifs du TP**

Nous allons effectuer des tâches d’intégration de données avec Talend Open Studio. Notre support sera une base de données de location de films. La BD englobe les données des films, des réalisateurs, des acteurs ainsi que des clients.

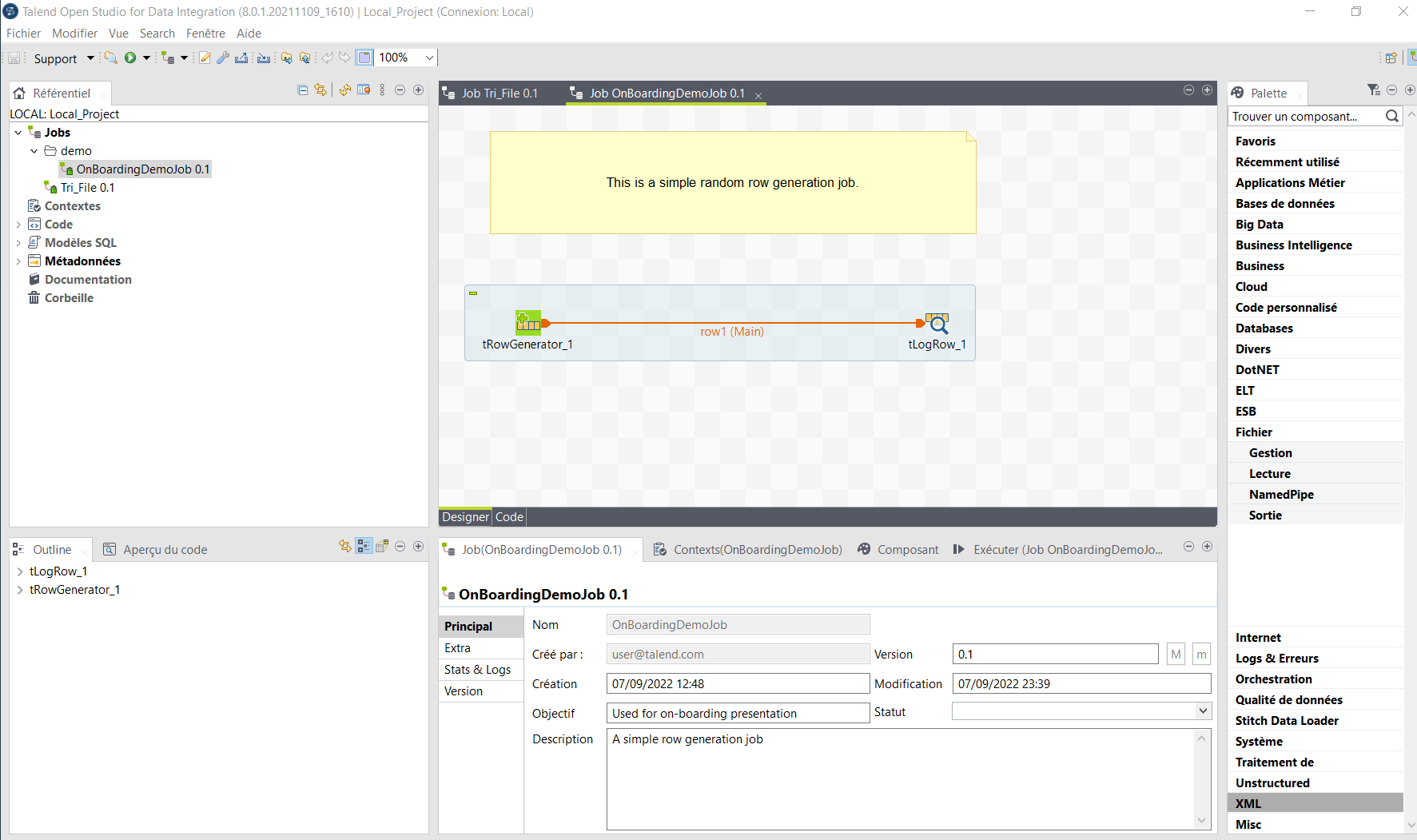
Nous allons entamer le processus ETL avec *Talend Studio Data Integration* dans le but de:

* Produire des données propres à partir de données brutes et qui peuvent être exploitées efficacement dans la phase d’analyse.
* Centraliser et homogénéiser les ressources de données
* Enregistrer les données extraites et transformées dans un *ED.*

**Installation et démarrage de Talend open Studio fo Data Integration V 8.0.1**

1. Téléchargez Talend en se référant au site officiel: [www.talend.com/download](http://www.talend.com/download)
2. Téléchargez et installez le jdk 11 ou plus pour que la version 8.0.1 puisse fonctionner correctement
3. Installez Talend sous un dossier nommé «Talend» créer sous C:
4. Une fois installé essayez de le consulter et de lancer l’outil.
5. Ouvrir votre outil et commencez à le découvrir.

**Description de l’interface de l’outil**



4

3

2

1

|  |  |
| --- | --- |
| 1 | **Référentiel:**  L’endroit où Talend Studio représente les données relatifs aux éléments techniques des différents jobs |
| 2 | **Espace de modélisation graphique:**  Composé de de deux onglets:  - *Designer*: affiche le job graphiquement  - *Code*: montre le code généré et identifie les erreurs possibles |
| 3 | **Onglet de configuration**:  Affiche les propriétés des composants, la console d’exécution, les erreurs, le contexte d’un Job, etc. |
| 4 | **Palette:**  Montre les différents *composants techniques* à utiliser pour construire un Job (groupés en catégories) |

Avant d’entamer nos différentes activités nous allons commencer par la création d’un nouveau Project TOS Data Integration en suivant les étapes suivantes:

* Ouvrir l'assistant de TOS for Data Integration en double-cliquant sur le raccourci du logiciel
* Sélectionnez «Créer un nouveau projet» puis «Ok»
* Saisissez le nom du projet puis cliquez sur «Terminer»
* Fermez l'onglet d'accueil en cliquant sur «Start now !»

### Activité 1

Dans cette activité, vous allez trier la table des «movies» par «**gender**» puis par «**year**».

**Tri des Movies par «gender» :**

1. Créez un nouveau **Job** intitulé «**tri\_movies\_gender**» à partir du référentiel de Talend
2. Créez une métadonnée pour le fichier CSV

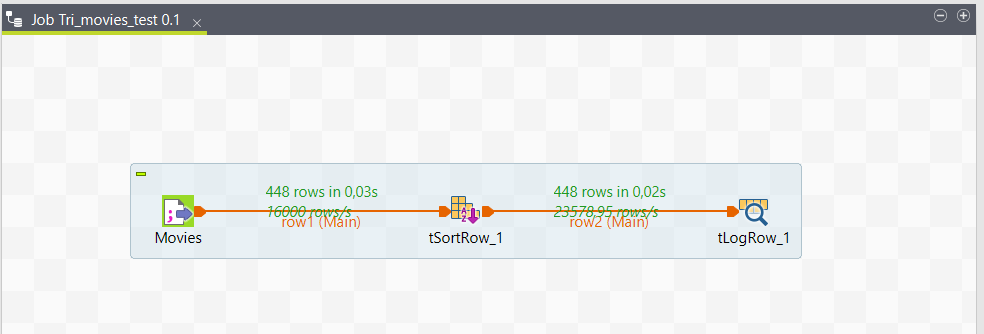
* Toujours dans le référentiel, développez la catégorie «**Métadonnées**»
* Clic droit sur «**Fichier délimité**», puis «**Créer un fichier délimité**»
* Dans l'assistant qui s'ouvre, entrez «**Movies**» dans le champ nom, puis cliquez sur «**Suivant**»
* Dans la fenêtre suivante, cliquez sur «**Parcourir**» et sélectionnez le fichier «**movies\_lim\_2023.csv**» à l'endroit approprié; sélectionnez également le bon format de votre ordinateur (son OS), puis faites «**Next**»
* En bas de l'assistant, dans la zone «**Aperçu**», attendez que le premier aperçu soit terminé, puis cliquez sur «**définir la ligne d'en-tête comme noms des colonnes**» puis sur «**Rafraîchir l'aperçu**», puis sur «**Next**»
* Dans la zone «**description du schéma**» cochez la case «**clé**» pour la première colonne (id\_movie) et décochez la case «**Nullable**» (pour définir la clé primaire), puis «**Finish**»

1. Glissez le fichier délimité «**Movies**» dans votre job vierge: un composant «**tFileInputDelimited**» est alors automatiquement créé dans votre job
2. Triez les colonnes du fichier délimité en se référant au composant «**tSortRow**»

* Dans la palette de composants à droite, trouvez le composant «**tSortRow**», glissez-le et déposez-le dans le job designer
* Affichez le résultat avec le composant «**tLogRow**» pour effectuer un simple test (affichage des résultats dans la console)
* Reliez les 3 composants (Movies → tSortRow → tLogRow) : créez donc 2 liens en cliquant droit sur le composant de départ, puis glissez ce lien jusqu'au composant d'arrivée, pour chaque lien. Nommez ces lien
* Double-clique sur «**tSortRow**» pour configurer ce composant et afficher les paramètres de transformation:

- cliquez sur (+) en bas du panneau de configuration, puis modifiez la valeur de la colonne du «shéma» en sélectionnant «**gender**» pour faire le tri par genre et par ordre alphabétique (**apha**) descendant (**desc**)

* Exécutez le job dans la vue Exécution (clic sur la flèche verte)
* Votre Job rassemble alors à ceci:



1. Triez le même fichier délimité par ordre ascendant (**asc**) et afficher le résultat dans la console

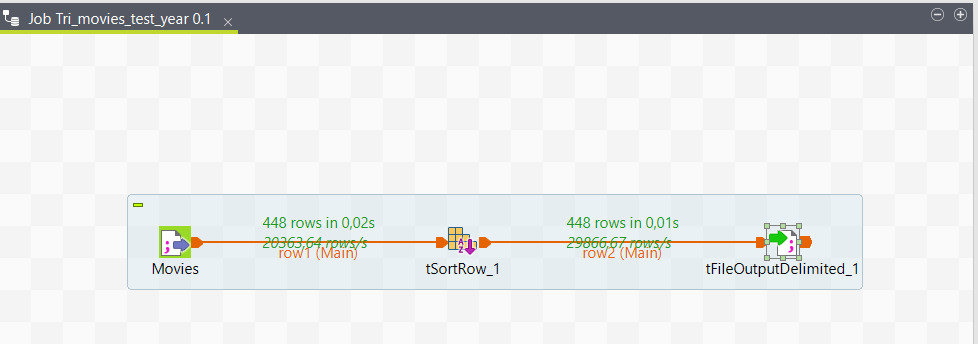
**Tri des Movies par «year» :**

Dans une deuxième partie de cette activité, vous allez procéder de même afin de trier les Movies par «**year**», bien que les résultats seront affichés dans un fichier CSV:

1. Créez un nouveau job intitulé «**tri\_movies\_year**»
2. Glissez le fichier délimité «**Movies**» dans votre job vierge
3. Glissez à partir de la palette à droite le composant «**tSortRow**»
4. Ajoutez le composant de sortie «**tFileOutputDelimited**» à votre zone graphique
5. Reliez entre les 3 composants
6. Configurez votre composant «**tSortRow**» en assurant le trie des Movies par «**year**» et par ordre croissant
7. Configurez votre fichier de sortie «**tFileOutputDelimited**»:

* Double-clique sur le composant afin de définir ses «**Paramètres simples**»  dans l'onglet **Composant**
* À côté du champ «**Nom de fichier**» cliquez sur le bouton **[...]** afin de parcourir votre système jusqu'à l'emplacement où vous souhaitez sauvegarder le fichier de sortie, puis saisissez le nom «**out\_tri\_year.csv**» pour le fichier
* Cochez la case «**Inclure l’en-tête»**  afin de récupérer le nom des colonnes du schéma ainsi que les données

1. Exécutez le job et vérifiez que le CSV a bien été crée
2. Votre job ressemble alors à ceci :



**Activité 2**

Dans cette activité, vous allez remplir les cases vides dans la colonne «**gender**» de la table «**Movies**» par le mot «**Empty**» et afficher les résultats dans deux sorties différentes: la console de Talend et un fichier CSV

1. Créez un nouveau job intitulé «**replace\_gender**»
2. Glissez le fichier délimité «**Movies**» dans votre job vierge
3. Glissez à partir de la palette à droite, le composant «**tMap**»
4. Ajoutez les composants de sortie: «**tLogRow**» et «**tFileOutputDelimited**» à votre zone graphique
5. Reliez entre les 4 composants:

* Entre «**tMap**» et «**tFileOutputDelimited**» le système demande l’affectation d’un nouveau nom de sortie: écrivez alors «**out\_file**»
* Entre «**tMap**» et «**tLogRow**» le système demande aussi l’affectation d’un nouveau nom de sortie: écrivez alors «**out\_console**»

1. Une erreur apparaît sur «**tMap**», double clique ce composant pour afficher les paramètres de transformation:

* Glissez-déposez tous les attributs de la table «**Movies**» vers la table de sortie «**out\_file**»
* Glissez-déposez l’attribut «**gender**» de la table «**Movies**» vers la table de sortie «**out\_console**»
* Mettez votre souris sur l'attribut «**gender**» de la table de sortie, "**...**" s'affiche, cliquez dessus
* Dans la partie «**Expression**», saisissez l’appel d’une fonction Java que vous allez créez vous même:

routines.Replace.replace\_gender(row1.gender)

1. Création de votre fonction Java:

* Créez une **routine** Talend : dans le référentiel, sous l'arborescence «**Code**», cliquez-droit sur «**Routines globales**» et sélectionnez «**créer une routine**»
* Nommez votre routine «**Replace**» et cliquez sur «**Finish**»
* Une fenêtre s'affiche avec une classe java portant le même nom que votre routine
* Ajoutez à votre classe une nouvelle fonction nommée «**replace\_gender**» qui prend en input une chaîne de caractères, teste si cette chaîne est vide ou non. Si la condition est vérifiée, la fonction remplace la chaîne vide par le mot «Empty» si non elle retourne la chaîne trouvée

**public** **static** String replace\_gender (String s) {

**if** (s.equals(""))

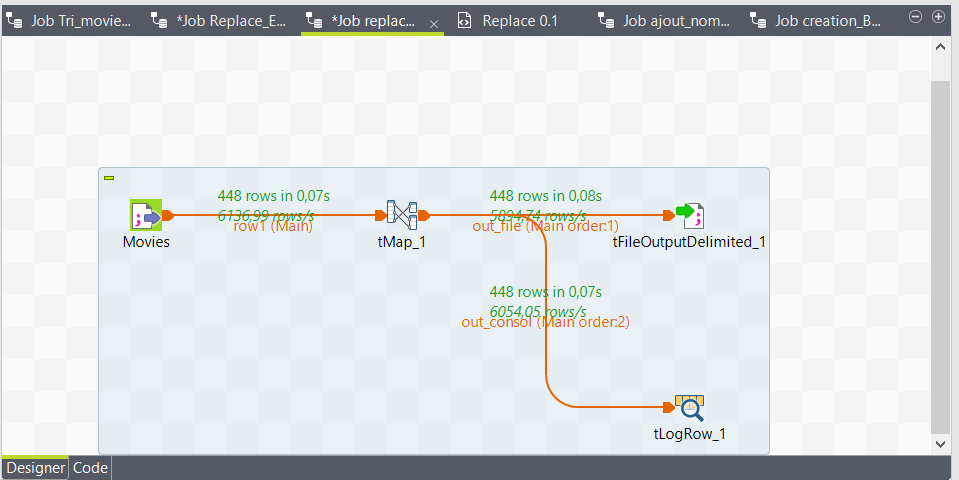
**return** "Empty";

**else**

**return** s;

}

* Sauvegardez votre routine et retournez dans le job initial
* Configurez le fichier «**tFileOutputDelimited**»: nommez le fichier de sortie «**out\_replace\_empty.csv**» et cochez la case «**Inclure l’en-tête**» afin de récupérer le nom des colonnes du schéma ainsi que les données
* Exécutez le job et vérifiez que le CSV a bien été crée
* Votre job ressemble alors à ceci :



**Activité 3**

Dans cette activité, vous allez transformer les identifiants des réalisateurs dans la table «Directors» pour qu'ils soient les mêmes que ceux renseignés dans le fichier «Movies»: enlevez les 0 devant les numéros de directors inférieurs à 10

1. Céez un nouveau job intitulé «**replace\_id\_director**»
2. Créez une métadonnée pour le fichier CSV «**Directors\_2023.csv**»
3. Glissez le fichier délimité «**Directors**» dans votre job vierge
4. Glissez à partir de la palette à droite les composants «**tMap**», «**tLogRow**» et «**tFileOutputDelimited**» à votre zone graphique
5. Reliez entre les 4 composants
6. Configurez le composant «**tLogRow**»: dans la zone «**Mode**» sélectionnez l’option «**Vertical (chaque ligne est une liste clé/valeur)**» pour une meilleure lisibilité des informations sur la console
7. Configurer la transformation de «**tMap**»:

* Glissez-déposez tous les attributs de la table de directors vers les tables de sorties
* Mettez votre souris sur l'attribut «**id\_Director**» des tables de sorties, "..." s'affiche, cliquez dessus
* Dans la partie «**Expression**» saisissez la fonction Java: Integer.parseInt(xxxx.id\_Director) , avec xxxx le nom de votre table d’entrée
* Puis changez, dans l’éditeur de schéma de chacune des deux tables de sorties (en bas à droite), le type de cet attribut de «**String**» à «**int | Integer**», et cliquez sur «**OK**»
* Configurez le fichier «**tFileOutputDelimited**»: nommez le fichier de sortie «**out\_id\_director.csv**» et cochez la case «**Inclure l’en-tête**».
* Exécutez le job dans la vue Exécution
* Votre job ressemble alors à ceci :

