Talend

Intégration des données d’un fichier Excel pas à pas

[Présentation 1](#_Toc279849973)

[Mise en œuvre 1](#_Toc279849974)

[Intégration des données du fichier Années Darties 2](#_Toc279849975)

[Guide pas à pas 2](#_Toc279849976)

[Premier lancement de Talend 2](#_Toc279849977)

[Paramétrage du fichier source et de la base de données 3](#_Toc279849978)

[1er job : intégration des continents 5](#_Toc279849979)

# Présentation

Talend Open Studio est un logiciel d’intégration de données open source, développé par la société parisienne homonyme. Il a été distribué pour la première fois en octobre 2006. Une extension payante est aussi proposée par l’éditeur : Talend Integration Suite, qui inclut des fonctionnalités supplémentaires et un support technique.

Talend est basé sur Eclipse RCP, et possède donc un certain nombre de similitudes avec Eclipse au niveau de l’interface.

Dans ce document, nous utiliserons Talend Open Studio v4.1.1.

# Mise en œuvre

La première chose que nous avons essayé de mettre en pratique est l’intégration de continents dans la base de données depuis le fichier Excel. Nous avons pour cela ajouté un job contenant un composant tExcelInput configuré pour lire la première feuille du fichier Annees\_Darties, et un composant tOracleOutput configuré pour écrire dans la base de données Oracle. Nous avons ensuite utilisé un tMap afin d’effectuer le mapping entre les données du fichier et de la base.

La première difficulté a été d’utiliser les séquences Oracle pour la valeur des clés primaires lors de l’insertion. Après quelques recherches, nous avons pu trouver une solution qui permettait d’entrer directement du code SQL grâce à l’option « colonnes supplémentaires » de Talend. La nécessité d’utiliser cette fonctionnalité pour arriver à nos fins ne nous a pas paru très évidente, mais cela avait le mérite de très bien fonctionner.

Notre première stratégie a été de créer un seul job pour effectuer toutes les initialisations : un tInputExcel pour lire toutes les données, plusieurs tOracleInput pour effectuer des jointures avec les données existantes, plusieurs tOracleOutput correspondant aux différentes tables dans lesquelles nous avions à insérer des données, et enfin un tMap central permettant d’effectuer tout le mapping nécessaire.

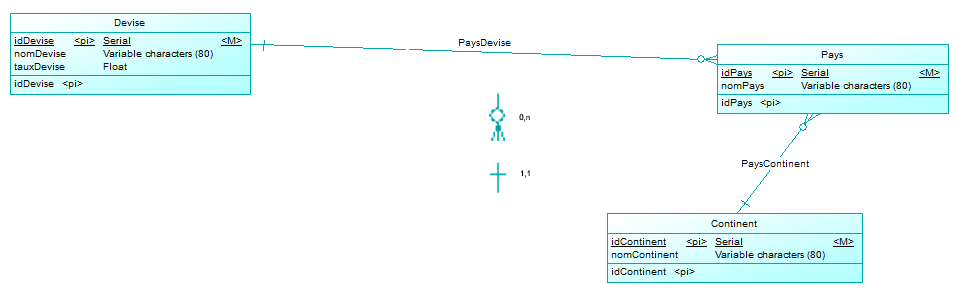
Si cette solution fonctionnait très bien dans les cas simples, elle posait un problème majeur dans les situations où il fallait à la fois insérer plusieurs données dépendant les unes des autres. En effet, les données existantes n’étaient lues qu’une seule fois, au tout début ; les autres données insérées au cours du job n’étaient pas prises en compte, ce qui générait par la suite des erreurs de contraintes d’intégrité. La solution a été de diviser les initialisations en plusieurs jobs, et ensuite de créer un dernier job chargé d’effectuer tous les autres dans le bon ordre.

Une autre difficulté est apparue par la suite, quand il s’est agi d’insérer les données de la table Faits. Chaque ligne correspondait aux données de toute l’année, objectifs et résultats, pour une ville et un article donné. La base de données étant structurée d’une manière complètement différente, nous arrivions donc à la nécessité d’insérer 12 tuples pour chaque ligne du fichier Excel. Cela nécessitait donc un autre mécanisme plus complexe que du simple mapping ; ceci combiné aux autres difficultés que posaient les règles de nommage employées dans le fichier Darties, nous avons plutôt choisi de modifier la structure du fichier Excel afin d’avoir une architecture plus exploitable tout en restant lisible et cohérente. Ces modifications sont expliquées dans le document de suivi du fichier Darties.

# Intégration des données du fichier Années Darties

Nous allons ici présenter un exemple d’intégration de données issues du fichier Excel Annees Darties.xls. Nous nous intéresserons à l’importation des continents.

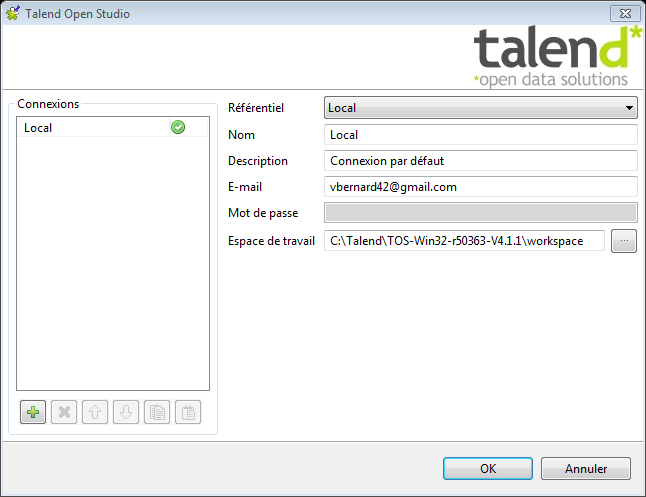
Voici la partie correspondante du MCD :



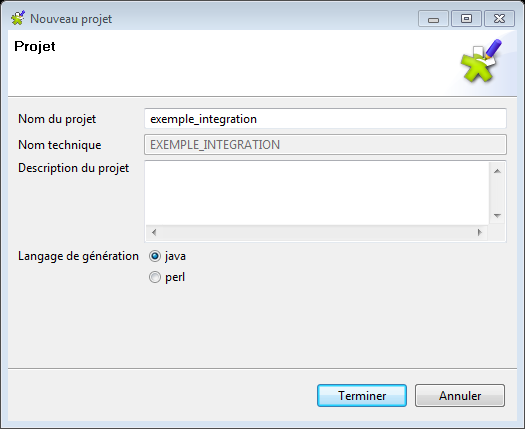
On note qu’un pays appartient à un continent et possède une devise ; les deux éléments desquels il dépend devront donc être insérés en premier afin de préserver l’intégrité de la base. Nous allons ici uniquement créer le job insérant les continents.

# Guide pas à pas

## Premier lancement de Talend

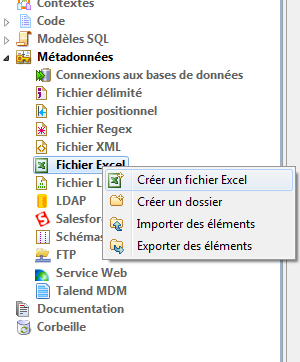


Au premier lancement de Talend, il sera demandé de créer un référentiel. Cette étape est rapide : il suffit de cliquer sur le bouton « … », d’entrer une adresse e-mail quelconque et de valider.



Nous choisissons ensuite de créer un nouveau projet en local que nous nommons exemple\_integration, et validons. Nous pouvons ensuite ouvrir le projet.

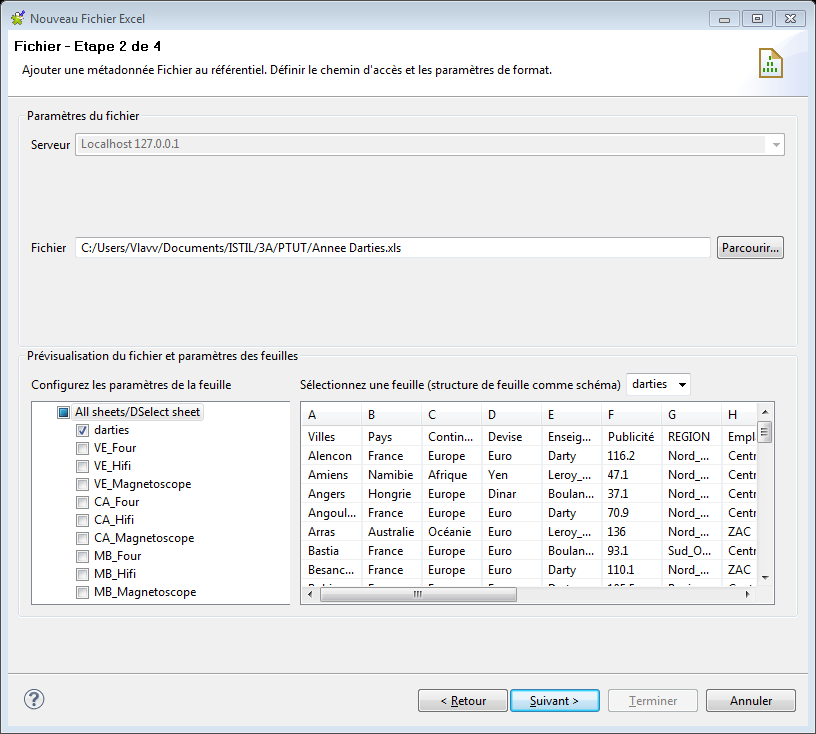
## Paramétrage du fichier source et de la base de données

Une fois le projet ouvert, l’onglet Bienvenue s’affiche, que nous pouvons fermer afin d’obtenir l’arborescence de notre référentiel.

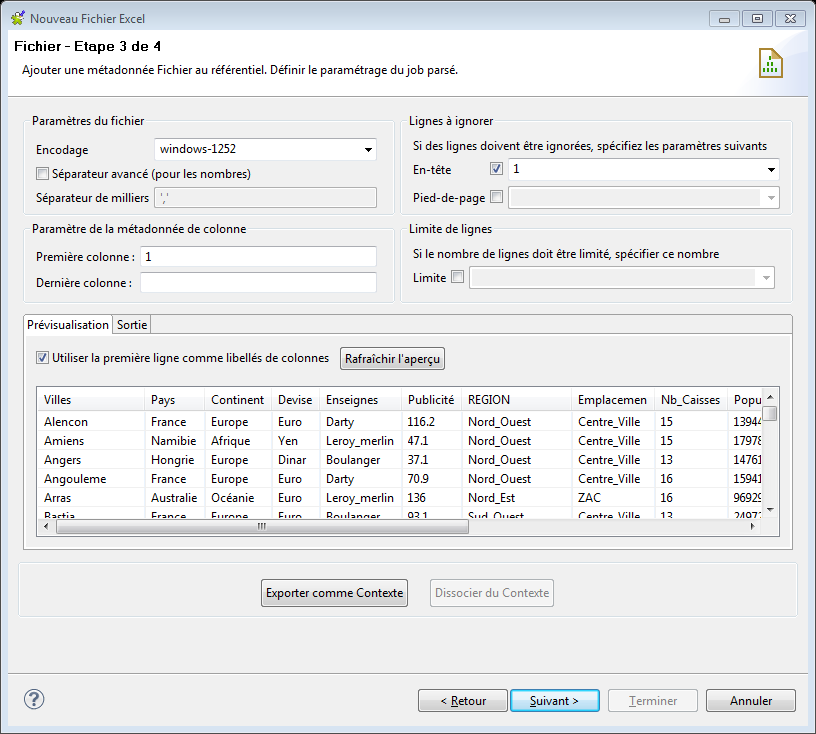
Nous allons d’abord déclarer notre fichier Excel, afin de pouvoir l’utiliser facilement avec les composants Talend par la suite.

L’opération s’effectue dans le menu contextuel de l’élément Fichier Excel de la catégorie Métadonnées.

Nous nommons le fichier annees\_darties, et cliquons sur Suivant.

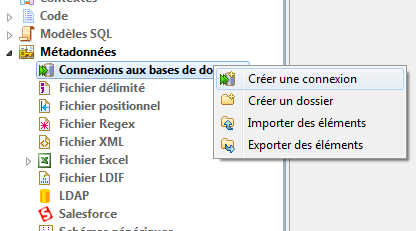


Nous indiquons ensuite l’emplacement du fichier, et cliquons sur Suivant à nouveau.

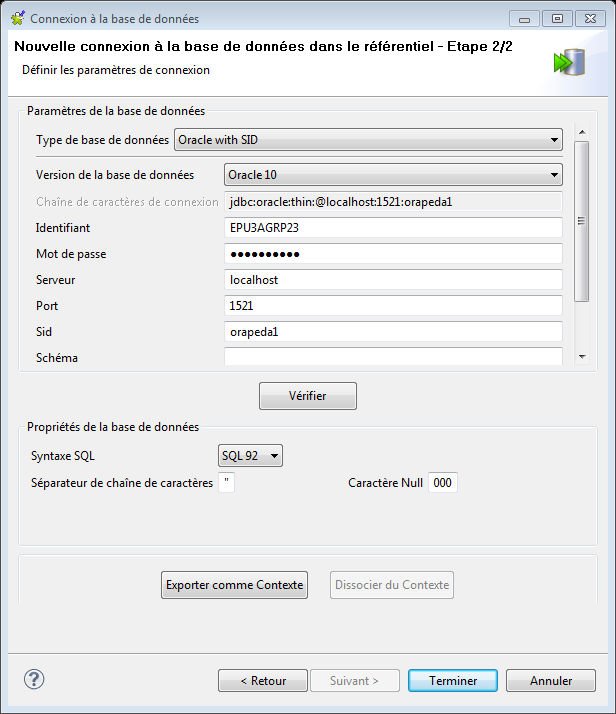


Nous précisons enfin que nous avons une ligne d’en-tête, et que l’encodage du fichier n’est pas UTF-8 mais windows-1252 (@#!).

Suivant, puis Terminer.

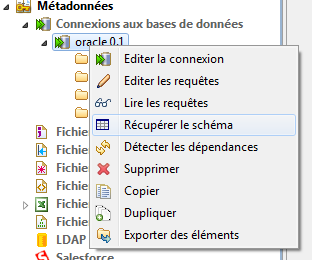


De la même façon, nous allons ajouter la connexion à notre base de données, que nous pouvons nommer oracle.

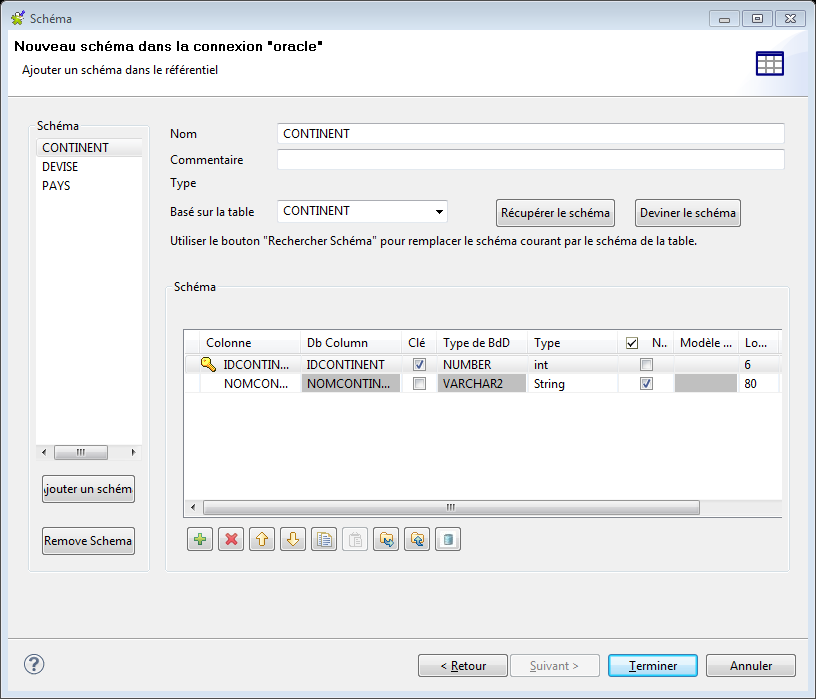


Nous entrons les paramètres de connexion de la base, et cliquons sur Terminer.

S’ensuit une étape qui peut être longue et douloureuse : la récupération du schéma de la base.



Allez chercher un bon bouquin, mettez de la musique douce, et cliquez ici.

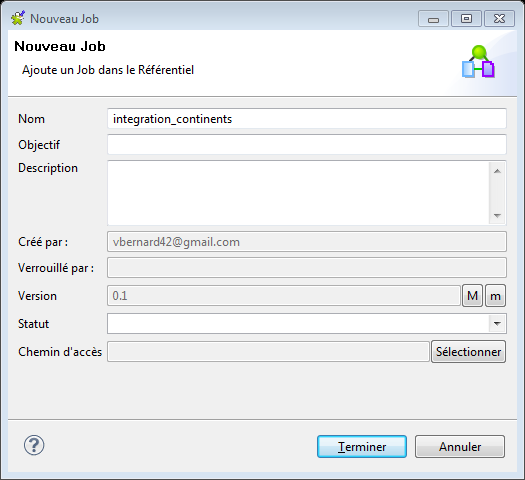


Nous décochons VIEW et SYNONYM, puis cliquons sur Suivant.

Nous cochons ensuite les tables dont nous avons besoin (CONTINENT, DEVISE et PAYS), puis attendons qu’il soit possible de cliquer sur Suivant.

Dans le dernier écran, nous allons modifier un peu le mapping des types afin d’obtenir une correspondance avec les données du fichier Excel. Pour chacune des trois tables, il faut changer les BigDecimal en int. Courage : il est possible d’effectuer cette opération en moins de deux heures et demie.

## 1er job : intégration des continents

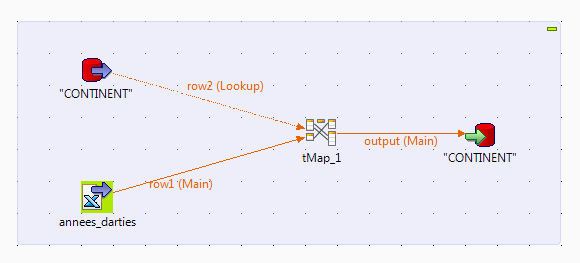


Nous créons un nouveau job que nous allons appeler integration\_continents.

Le job s’ajoute dans l’arborescence et s’ouvre directement dans le panneau principal.

Nous commençons par faire glisser notre fichier Excel depuis l’arborescence (dans Métadonnées) vers le panneau du job. Nous faisons de même pour notre table CONTINENT (Métadonnées/Connexions aux bases de données/Oracle/Schéma des tables), et nous précisons dans la fenêtre qui s’ouvre que nous voulons un tOracleInput. Nous ajoutons à nouveau la table CONTINENT, mais cette fois-ci en tant que tOracleOutput.

Enfin, nous ajoutons un composant tMap, disponible dans la section « Transformations » de la palette à droite.



Pour finir, nous relions tous nos composants avec des cliquer-déplacer du bouton droit de la souris, en commençant par le fichier Excel. Nous pouvons nommer le lien de sortie du tMap « output ».

### Configuration du tMap

Nous double-cliquons sur le tMap que nous venons d’ajouter. Nous obtenons une fenêtre permettant d’effectuer le mapping entre les composants d’entrée et de sortie.

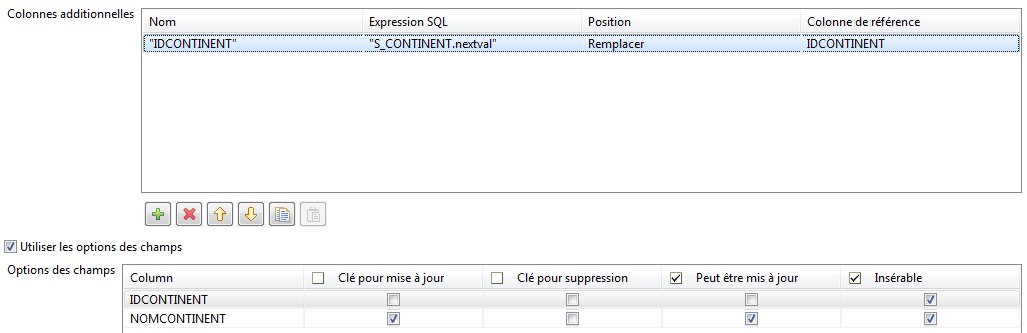
Dans la partie gauche, nous faisons un cliquer-déplacer du Continent du fichier Excel vers le NOMCONTINENT de la base de données. Cela permet de créer une jointure, et de récupérer l’identifiant du continent à partir de son nom (s’il existe déjà).

Nous faisons ensuite des glisser-déplacer de la même façon afin d’obtenir le schéma ci-dessous.

### Configuration du tOracleOutput

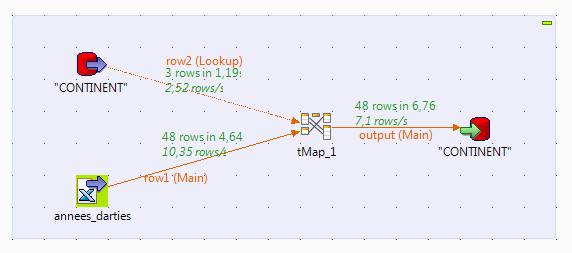
De retour dans le job, nous allons dans les propriétés du composant CONTINENT, et modifions le champ « Action sur les données » de « Insert » à « Insert ou update » dans la section « Paramètres simples ».

Dans la section « Paramètres avancés », nous cliquons sur le bouton **+** en dessous du tableau « colonnes additionnelles », et cochons la case « Utiliser les options des champs ». Nous remplissons les champs de la façon suivante :



Ceci permet d’utiliser la séquence Oracle pour générer de nouvelles clés.

### Exécution du job

Nous pouvons exécuter le job dans le panneau homonyme. Les continents s’insèrent dans la base s’ils n’existent pas déjà.