

# TALEND

## INTÉGRATION DES DONNÉES D'UN FICHIER EXCEL PAS À PAS

Présentation .....	1
Intégration des données du fichier Années Darties .....	1
Guide pas à pas .....	2
Premier lancement de Talend .....	2
Paramétrage du fichier source et de la base de données.....	2
1 <sup>er</sup> job : intégration des continents.....	4

### 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.

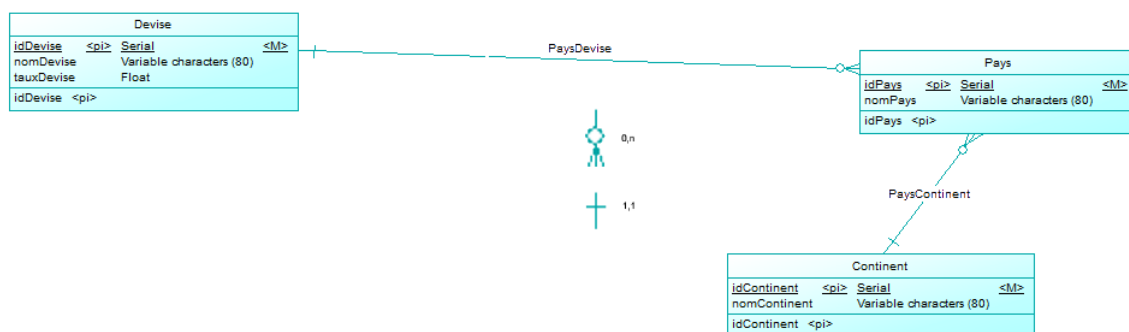
Talend est basé sur Eclipse RCP, et possède donc un certain nombre de similitudes avec Eclipse au niveau de l'interface.

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

### INTÉGRATION DES DONNÉES DU FICHIER ANNÉES DARTIES

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

Voici la partie correspondante du MCD :

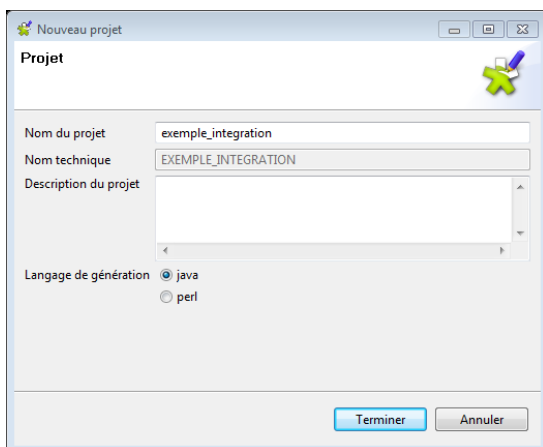
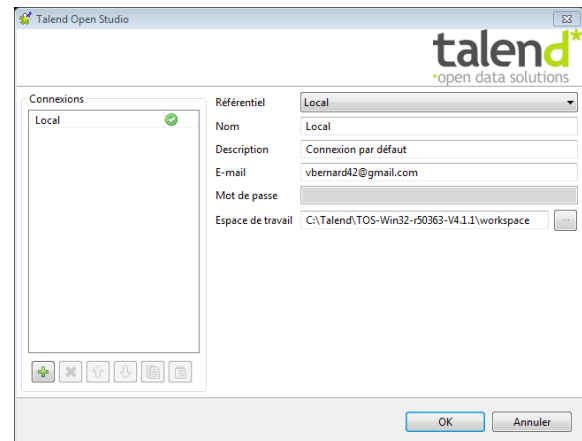


On note qu'un pays appartient à un continent et possède une devise ; les deux éléments desquels il dépend devront donc être insérés en premier afin de préserver l'intégrité de la base. Nous allons ici uniquement créer le job insérant les continents.

## 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.

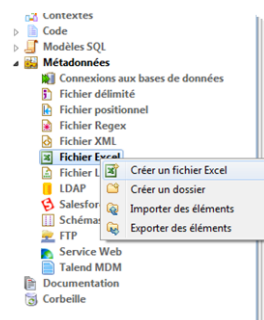
### PARAMÉTRAGE DU FICHIER SOURCE ET DE LA BASE DE DONNÉES

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*.



**Nouveau Fichier Excel**

**Fichier - Etape 2 de 4**

Ajouter une métadonnée Fichier au référentiel. Définir le chemin d'accès et les paramètres de format.

Paramètres du fichier

Serveur: Localhost 127.0.0.1

Fichier: C:/Users/Vlavo/Documents/ISTIL/3A/PTUT/Annee Daries.xls

Prévisualisation du fichier et paramètres des feuilles

Configurez les paramètres de la feuille

Sélectionnez une feuille (structure de feuille comme schéma) : daries

☒ All sheets/Select sheet:

- ☒ daries
- ☐ VE\_Four
- ☐ VE\_Hifi
- ☐ VE\_Magnetoscope
- ☐ CA\_Four
- ☐ CA\_Hifi
- ☐ CA\_Magnetoscope
- ☐ MB\_Four
- ☐ MB\_Hifi
- ☐ MB\_Magnetoscope

A	B	C	D	E	F	G	H
Villes	Pays	Contin...	Devis...	Enseig...	Publicité	REGION	Empl...
Alencon	France	Europe	Euro	Darty	116.2	Nord_...	Centr
Amiens	Namibie	Afrique	Yen	Leroy_...	47.1	Nord_...	Centr
Angers	Hongrie	Europe	Dinar	Boulan...	37.1	Nord_...	Centr
Angoul...	France	Europe	Euro	Darty	70.9	Nord_...	Centr
Arras	Australie	Océanie	Euro	Leroy_...	136	Nord_...	ZAC
Bastia	France	Europe	Euro	Boulan...	93.1	Sud_...	Centr
Besanc...	France	Europe	Euro	Darty	110.1	Nord_...	ZAC

< Retour Suivant > Terminer Annuler

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*.

**Nouveau Fichier Excel**

**Fichier - Etape 3 de 4**

Ajouter une métadonnée Fichier au référentiel. Définir le paramétrage du job parsé.

Paramètres du fichier

Encodage: windows-1252

☐ Séparateur avancé (pour les nombres)

Séparateur de milliers: ,

Paramètre de la métadonnée de colonne

Première colonne: 1

Dernière colonne:

Lignes à ignorer

Si des lignes doivent être ignorées, spécifiez les paramètres suivants

En-tête: ☒ 1

Pied-de-page:

Limite de lignes

Si le nombre de lignes doit être limité, spécifiez ce nombre

Limite:

Prévisualisation Sortie

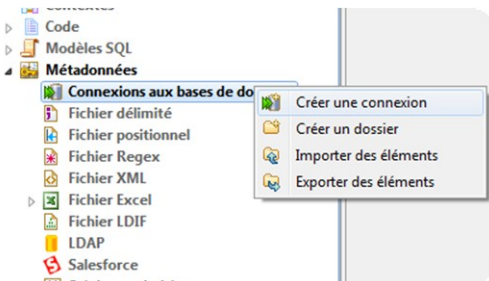
☒ Utiliser la première ligne comme libellés de colonnes

Rafraîchir l'aperçu

Villes	Pays	Continent	Devis...	Enseignes	Publicité	REGION	Emplacement	Nb_Cais...	Popu...
Alencon	France	Europe	Euro	Darty	116.2	Nord_Ouest	Centre_Ville	15	1394
Amiens	Namibie	Afrique	Yen	Leroy_merlin	47.1	Nord_Ouest	Centre_Ville	15	17976
Angers	Hongrie	Europe	Dinar	Boulanger	37.1	Nord_Ouest	Centre_Ville	13	14761
Angoulême	France	Europe	Euro	Darty	70.9	Nord_Ouest	Centre_Ville	16	15941
Arras	Australie	Océanie	Euro	Leroy_merlin	136	Nord_Est	ZAC	16	96925
Bastia	France	Europe	Euro	Boulanger	93.1	Sud_Ouest	Centre_Ville	13	24027

Exporter comme Contexte Dissocier du Contexte

< Retour Suivant > Terminer Annuler



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*.

**Connexion à la base de données**

**Nouvelle connexion à la base de données dans le référentiel - Etape 2/2**

Définir les paramètres de connexion

Paramètres de la base de données

Type de base de données: Oracle with SID

Versión de la base de données: Oracle 10

Chaîne de caractères de connexion: jdbc:oracle:thin:@localhost:1521:orapeda1

Identifiant: EPU3AGRP23

Mot de passe: .....

Serveur: localhost

Port: 1521

Sid: orapeda1

Schéma:

Vérifier

Propriétés de la base de données

Syntaxe SQL: SQL 92

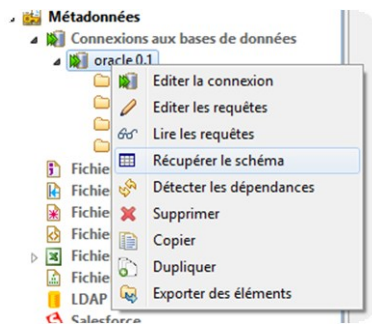
Séparateur de chaîne de caractères: "

Caractère Null: 000

Exporter comme Contexte Dissocier du Contexte

< Retour Suivant > Terminer Annuler

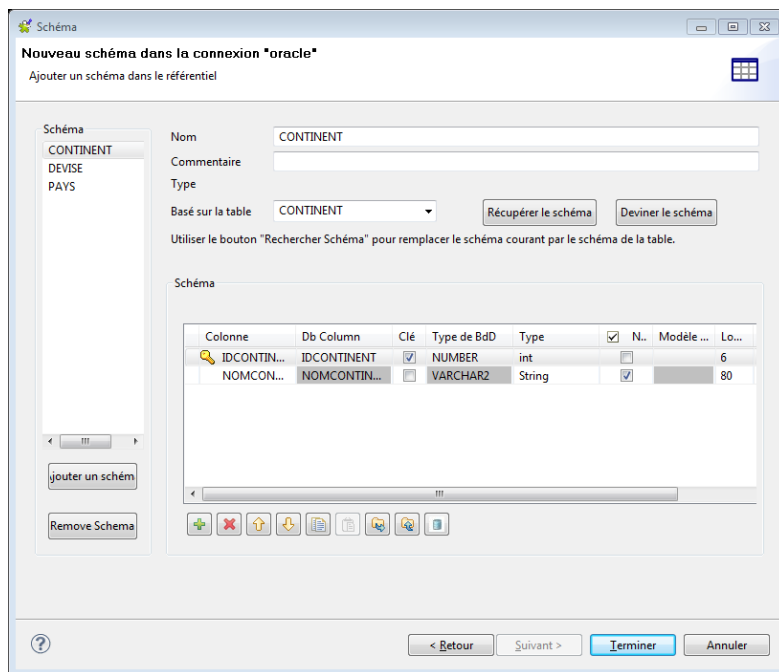
S'ensuit une étape qui peut être longue et douloureuse : la récupération du schéma de la base.



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

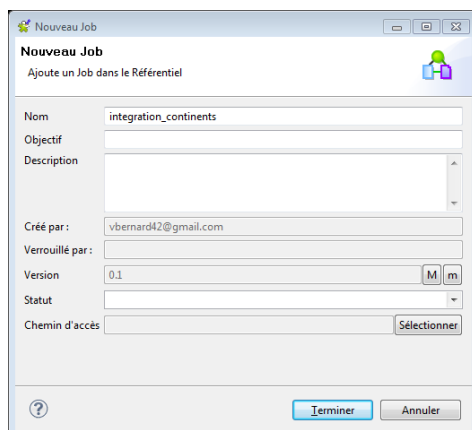
Nous décochons **VIEW** et **SYNONYM**, puis cliquons sur **Suivant**.

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.

## 1<sup>ER</sup> JOB : INTÉGRATION DES CONTINENTS



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

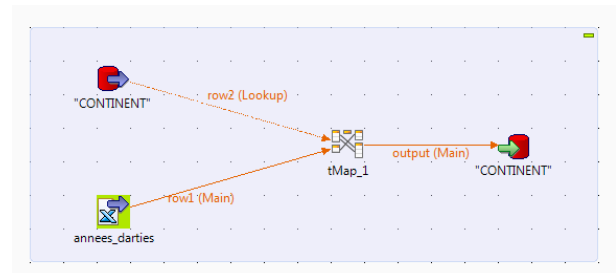
Le job s'ajoute dans l'arborescence et s'ouvre directement dans le panneau principal.

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 relierons 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 ».

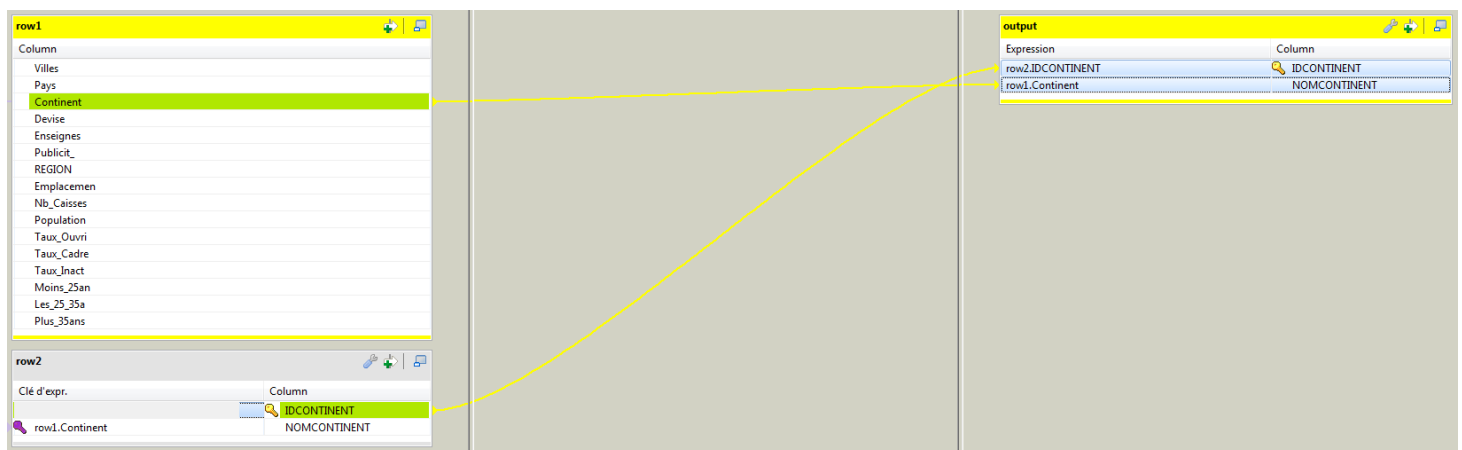


## CONFIGURATION DU TMAP

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 :

Colonnes additionnelles

Nom	Expression SQL	Position	Colonne de référence
IDCONTINENT	"S_CONTINENT.nextval"	Remplacer	IDCONTINENT

☒ Utiliser les options des champs

Options des champs

Column	<input type="checkbox"/> Clé pour mise à jour	<input type="checkbox"/> Clé pour suppression	<input checked="" type="checkbox"/> Peut être mis à jour	<input checked="" type="checkbox"/> Insérable
IDCONTINENT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
NOMCONTINENT	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>

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.

