TD « ENTREPOTS DE DONNEES »

Mise en œuvre de l'ETL TALEND – PARTIE 1

A - Trier un fichier : Créer un Job Design simple permettant de trier des données.

Ce tutoriel vous explique comment créer un Job permettant de lire les données d'un fichier délimité, écrire dans un fichier temporaire et remplacer le fichier original par ce fichier temporaire. Dans ce tutoriel, seuls les schémas de type "Built-in" ont été utilisés (les schémas ne sont pas stockés dans le référentiel mais sont relatifs à ce Job Design).

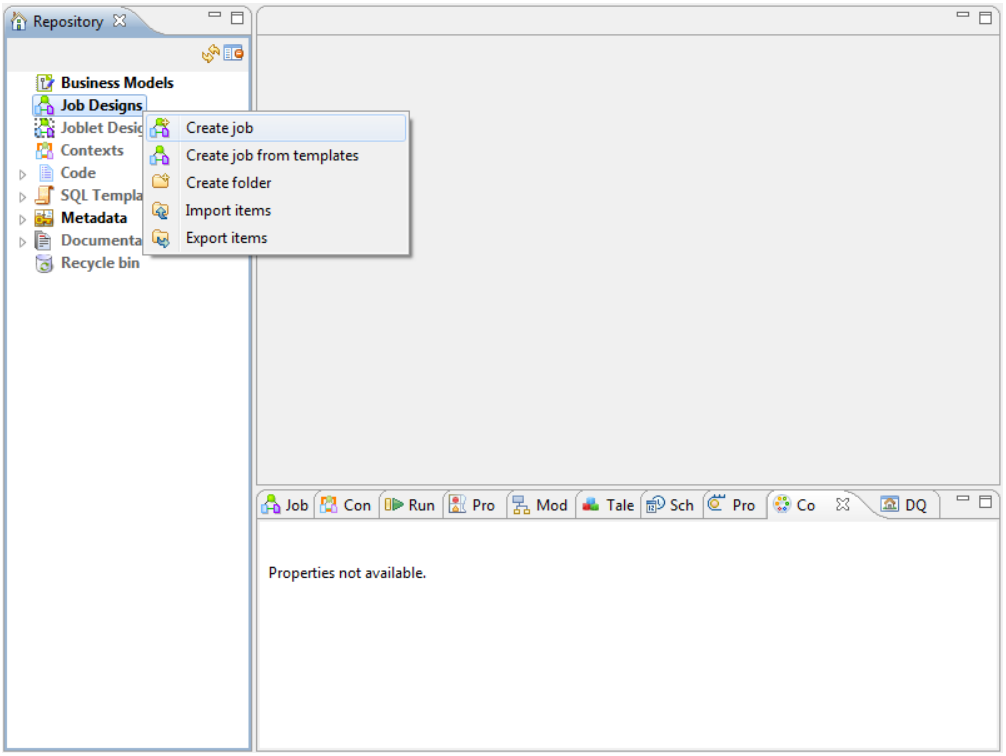
Prérequis : Pour suivre ce tutoriel, vous avez besoin d'extraire et d'importer le fichier customer.csv du fichier exampleFile.zip

Créer le Job Design

Dans le Repository situé à gauche de Talend Open Studio :

Pour créer un job, cliquez-droit sur **Job Designs**.

Dans le menu contextuel, cliquez sur **Create Job** pour ouvrir l'assistant **New Job**.



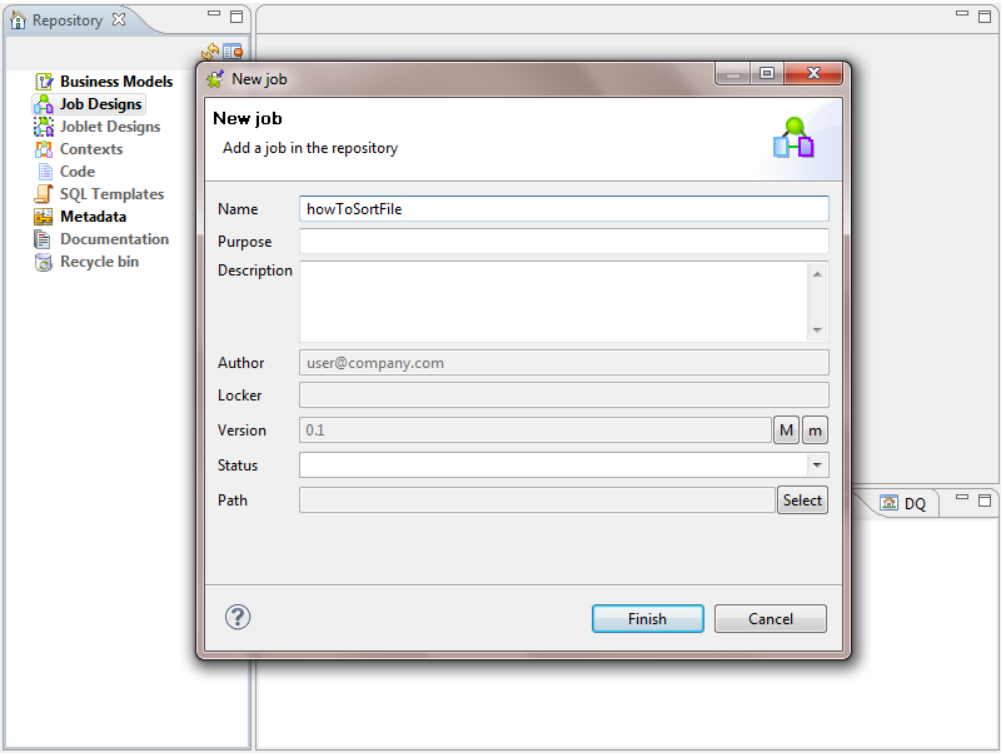
Dans l'assistant New Job :

Dans le champ **Name**, saisissez le nom du Job: *howToSortFile*.

Cliquez sur **Finish** pour fermer l'assistant et créer votre Job.

Le Job Designer présente alors un Job vierge.

Le champ **Name** ne doit pas contenir d'accents, de caractères spéciaux, d'espaces, ni débuter par un chiffre.

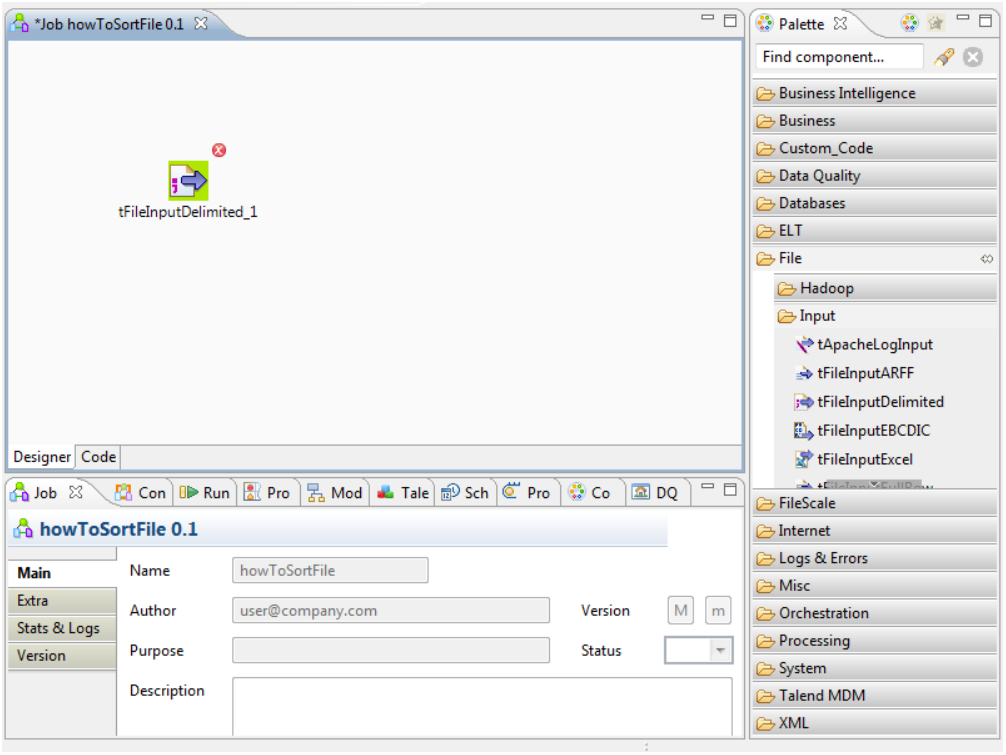


Paramétrer le connecteur de lecture de fichier délimité

Dans la Palette située à droite :

Pour ajouter le composant d'entrée, cliquez sur la famille **File** et sur la sous-famille **Input**.

Cliquez sur le composant **tFileInputDelimited** et déposez-le dans le Job Designer.



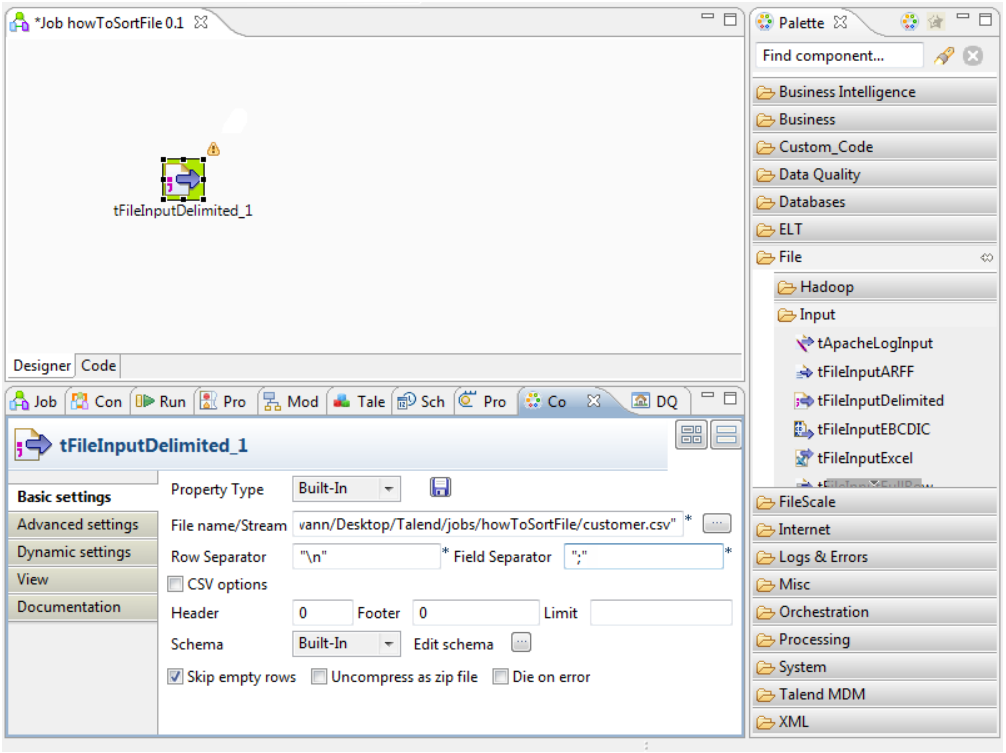
Dans le Job Designer :

Pour paramétrer les propriétés du **tFileInputDelimited**, double-cliquez sur le composant et la vue **Component** correspondante apparaît alors en bas de l'écran.

Dans la vue Component :

Pour spécifier le chemin d'accès au fichier *customer.csv*, cliquez sur le bouton **[...]** situé à coté du champ **File Name** et sélectionnez le fichier dans l'assistant qui s'ouvre alors.

Pour décrire la structure du fichier, cliquez sur le bouton **[...]** situé à coté du champ **Edit schema** pour ouvrir l'assistant "Schema of tFileInputDelimited

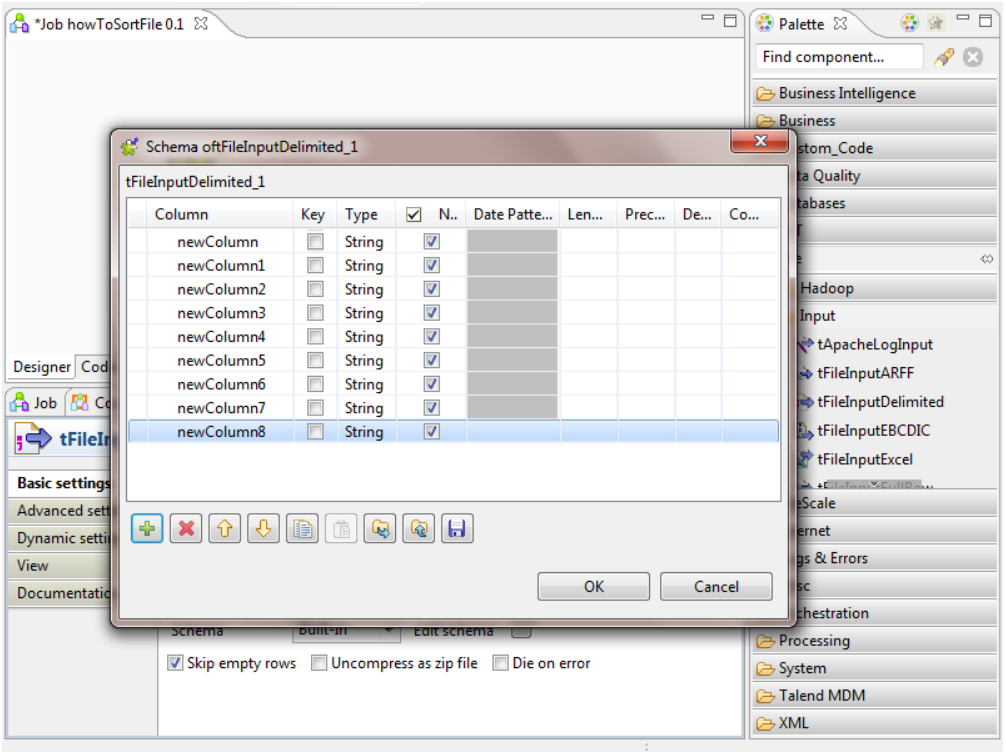


Configurer la structure du schéma du flux de données

Dans l'assistant Schema of tFileInputDelimited_1 :

Pour décrire les deux colonnes du fichier *customer*, cliquez neuf fois sur le bouton **[+]**. Cela ajoute neuf lignes au schéma correspondant aux colonnes du fichier.

Dans le cas de schémas présentant de nombreuses colonnes, l'utilisation des métadonnées est à privilégier.



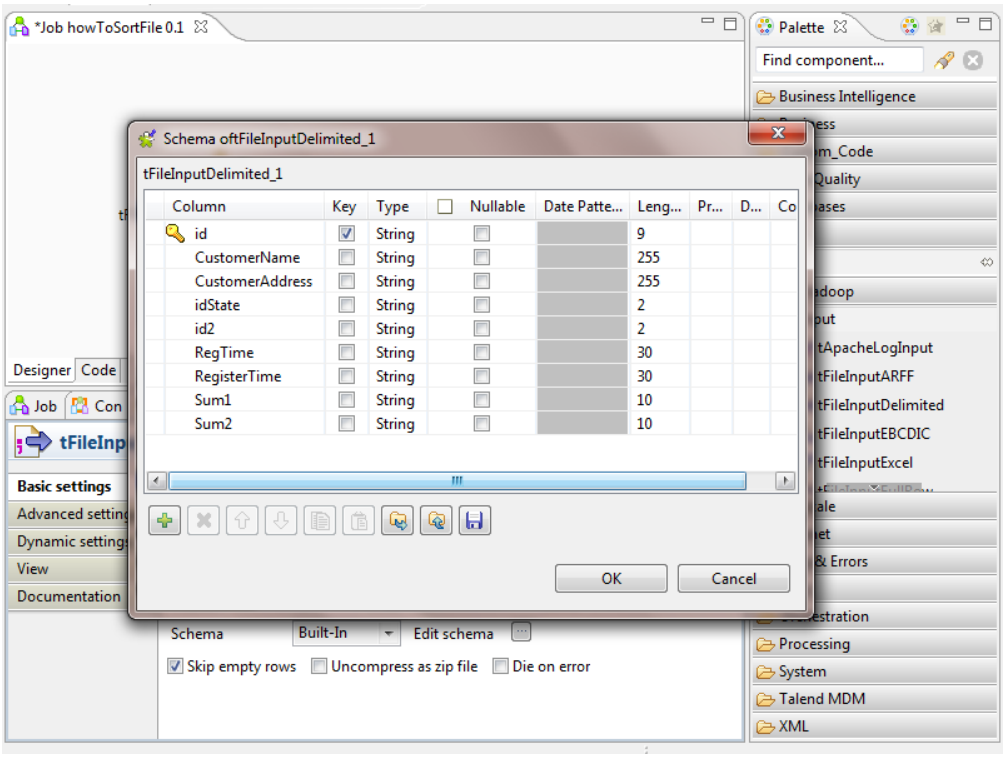
Dans l'assistant Schema of tFileInputDelimited_1 :

Dans la colonne **Column**, renommez chaque champ en fonction du nom des colonnes du fichier.

Dans la colonne **Type**, indiquez le type de champ pour chaque colonne.

Dans la colonne **Length**, renseignez la longueur pour chaque champ de votre schéma.

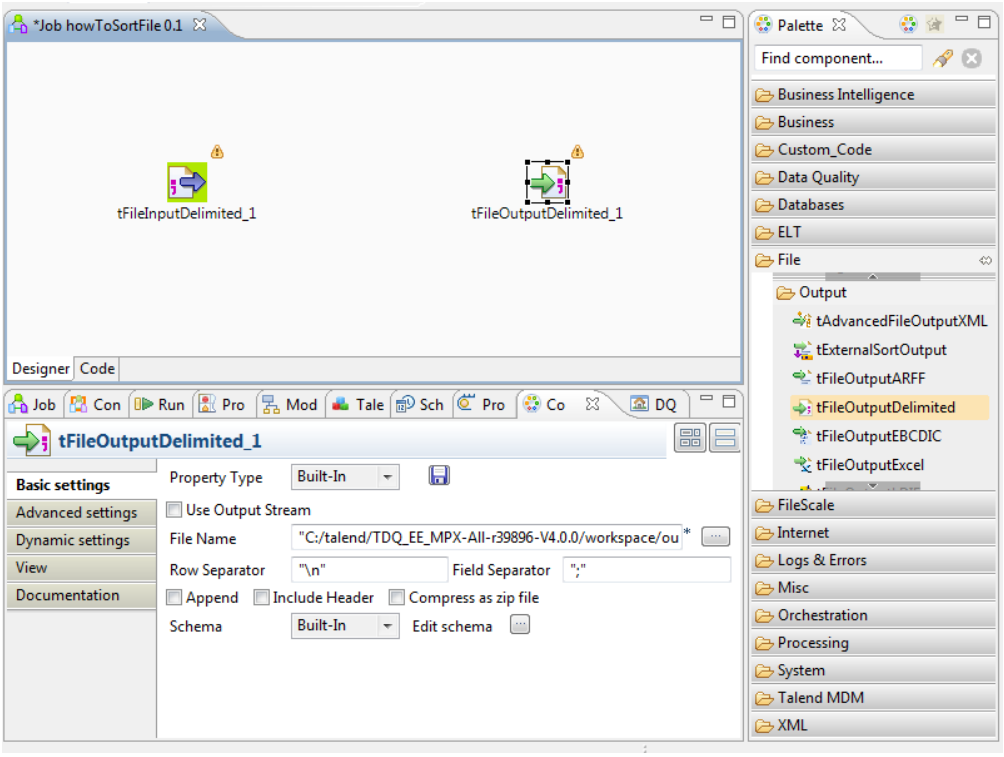
Cliquez sur **Ok** pour fermer l'assistant.



Dans la Palette située à droite :

Pour ajouter le composant de sortie, cliquez sur la sous-famille **Output**.

Cliquez sur le composant **tFileOutputDelimited** et déposez-le dans le Job Designer.



Dans le Job Designer :

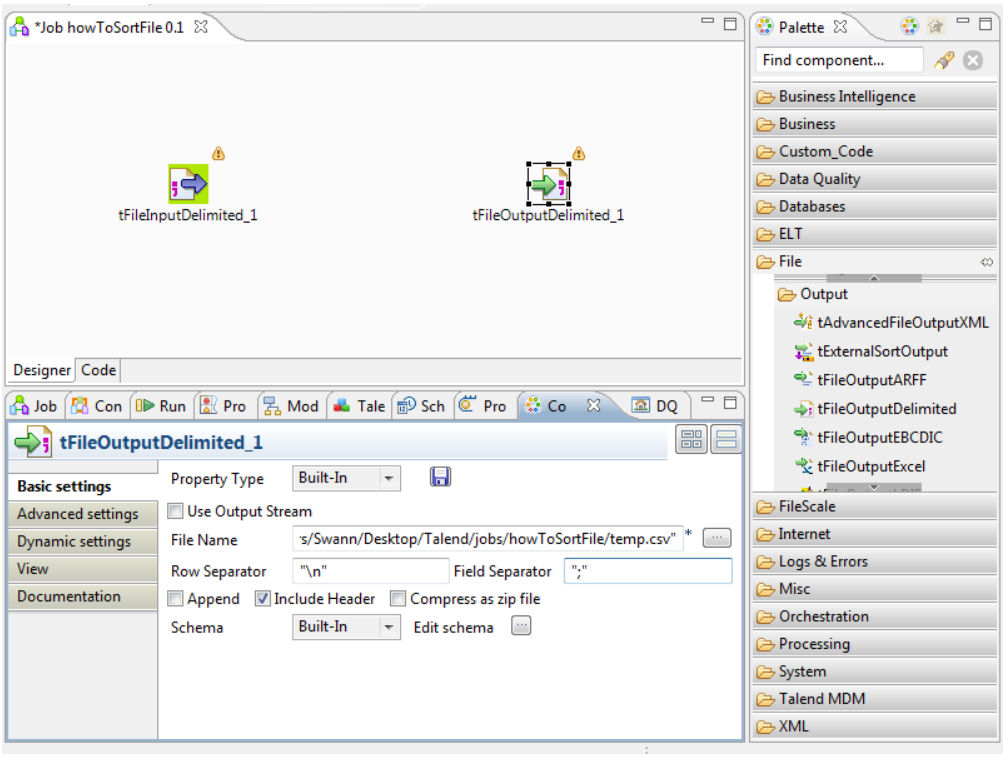
Pour paramétrer les propriétés du **tFileOutputDelimited**, double-cliquez dessus et la vue **Component** correspondante apparaît.

Dans la vue Component :

Pour spécifier le chemin du fichier qui sera créé, cliquez sur le bouton **[...]** situé à coté du champ **File Name**.

Grâce à l'assistant qui s'ouvre alors, définissez son chemin dans le même répertoire que le fichier *customer.csv* mais nommez-le *temp.csv*.

Cochez la case **Include Header** pour récupérer les noms des colonnes du fichier.



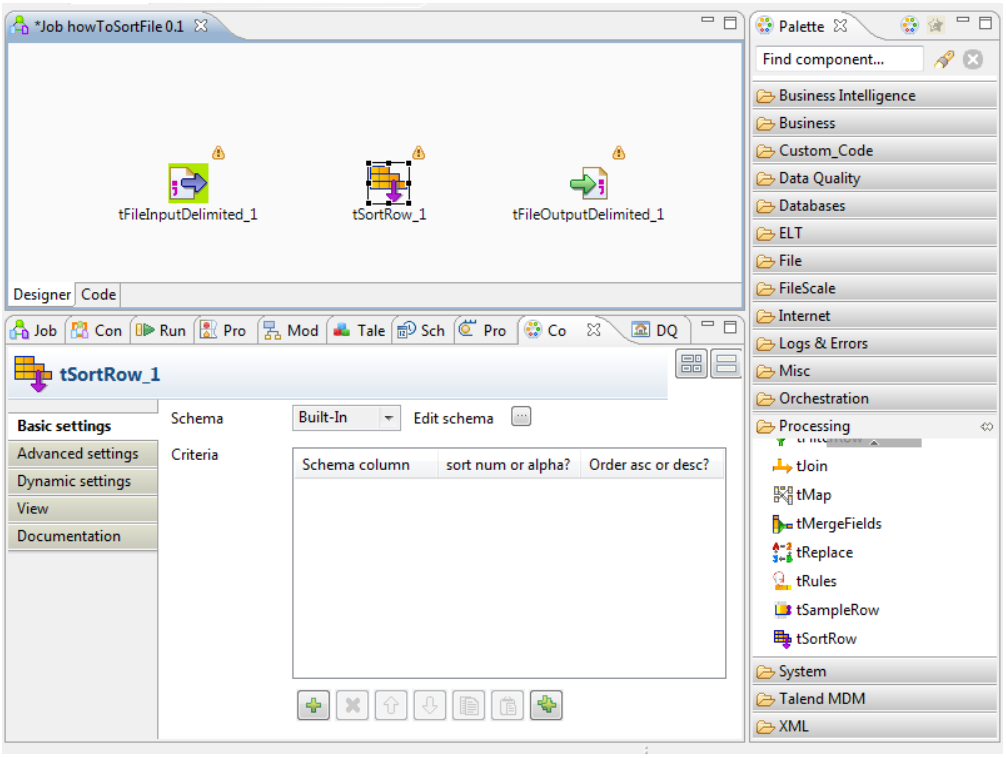
Définir le composant de transformation et relier les composants entre eux

---

Dans la Palette à droite :

Pour ajouter le composant qui va trier les données, cliquez sur la famille **Processing**.

Cliquez sur le composant **tSortRow** et déposez-le dans le Job Designer.

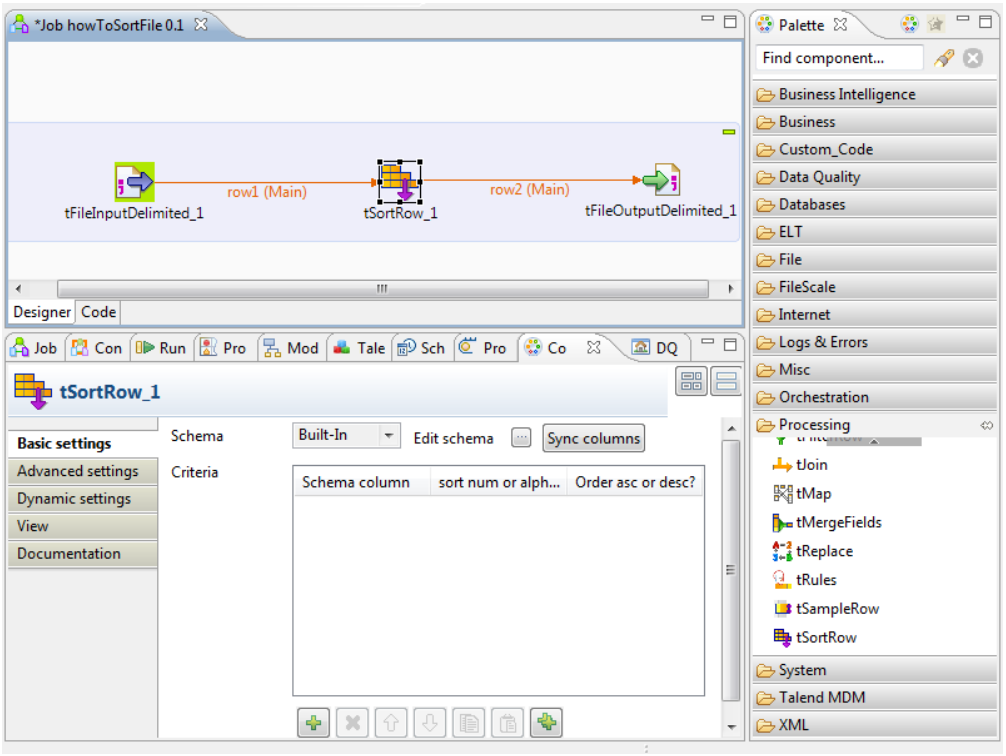


Dans le Job designer :

Pour relier les composants entre eux, cliquez-droit sur le **tFileInputDelimited** et, en gardant le bouton droit enfoncé, déplacez-vous jusqu'au **tSortRow** puis relâchez le bouton de la souris.

De la même manière, créez un lien du **tSortRow** vers le **tFileOutputDelimited**.

Vous pouvez aussi créer ce lien en cliquant droit sur le composant et en cliquant sur **Row** > **Main** dans le menu contextuel.



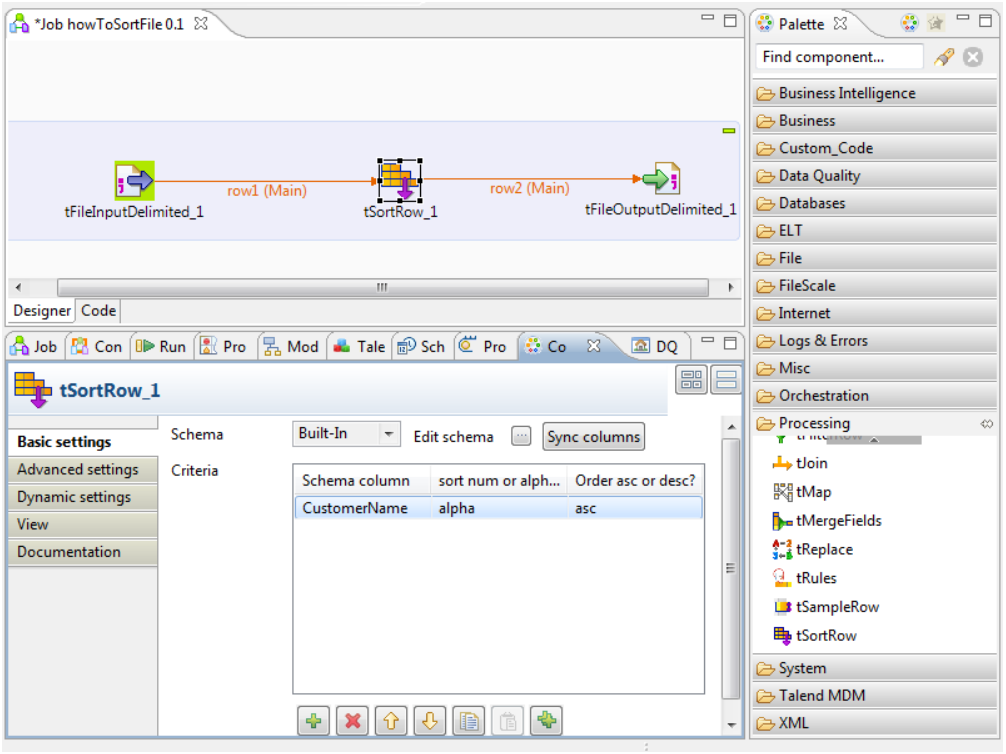
Dans le Job designer :

Pour paramétrer les propriétés du composant **tSortRow**, double-cliquez dessus et la vue **Component** correspondante apparaît.

Dans la vue Component :

Pour définir les critères de tri, cliquez sur le bouton **[+]** pour ajouter une ligne au tableau **Criteria**.

Sélectionnez la colonne que vous souhaitez trier comme indiqué dans la capture d'écran.



A ce stade, le Job va créer un nouveau fichier *temp.csv* contenant toutes les données triées.

L'objectif étant de trier le fichier original et non d'en créer un nouveau, il nous reste à remplacer le fichier original par ce nouveau fichier.

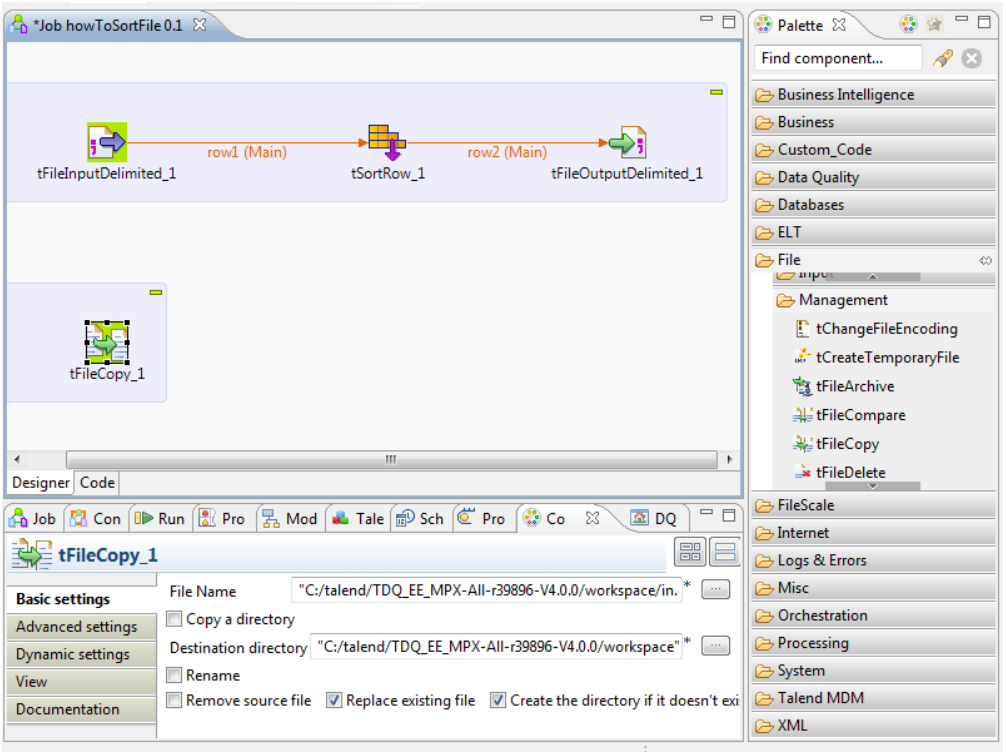
Définir le composant de manipulation de fichiers et le relier au sous-job précédent

---

Dans la Palette à droite :

Pour ajouter le composant permettant de remplacer le fichier original par le nouveau fichier trié, cliquez sur la famille **File** et sur le sous-famille **Management**.

Cliquez sur le composant **tFileCopy** et déposez-le dans le Job Designer, sous le composant **tFileInputDelimited**.



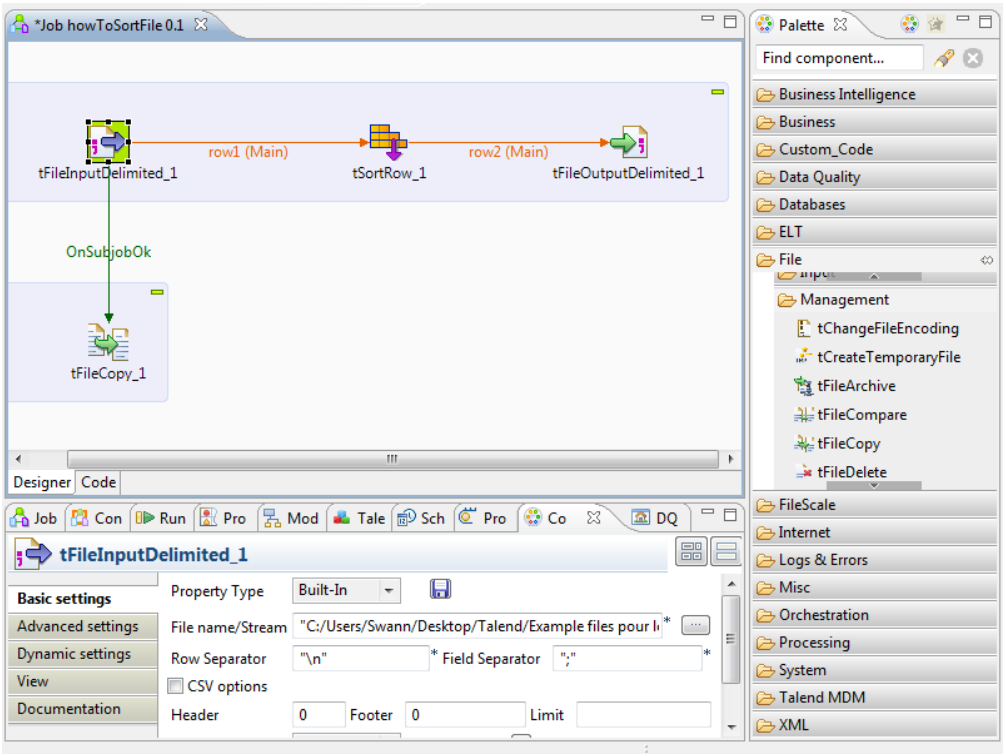
Dans le Job designer :

Pour relier notre premier sous-job au composant **tFileCopy**, cliquez-droit sur le **tFileInputDelimited** et sélectionnez **Trigger** > **OnSubjobOk** dans le menu contextuel.

Cliquez sur le **tFileCopy** pour dessiner le lien **OnSubjobOk**.

Dans le Job designer :

Pour paramétrer le composant **tFileCopy**, double-cliquez dessus et la vue **Component** correspondante apparaît.



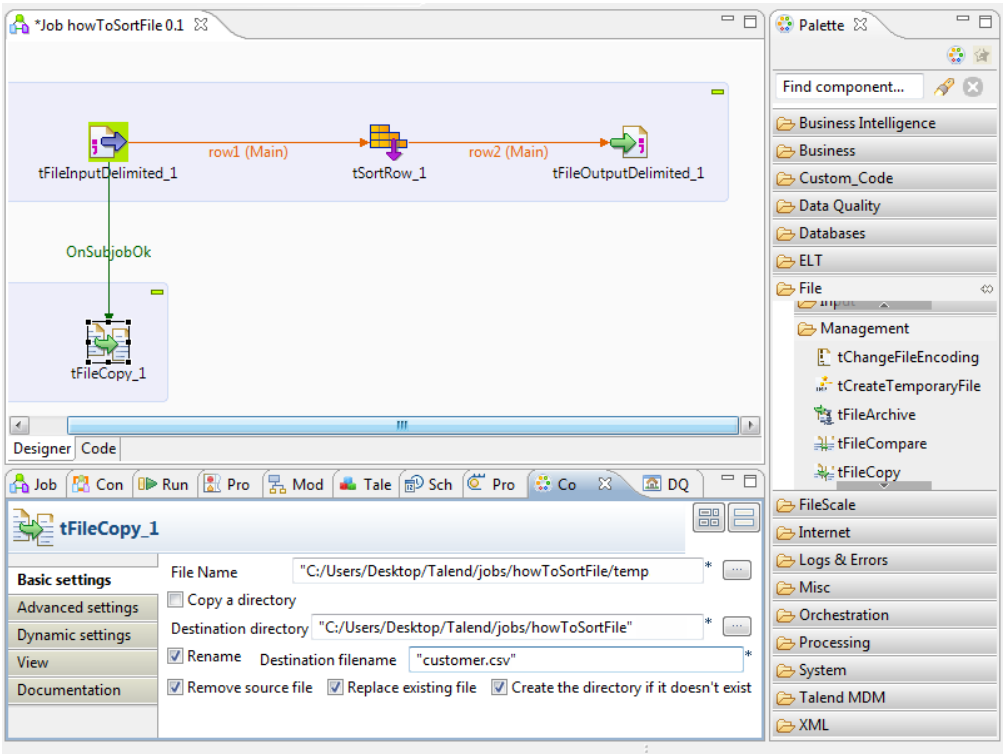
Dans la vue Component :

Pour copier le fichier *temp.csv* dont les données sont triées, cliquez sur le bouton **[...]** situé à coté du champ **File Name** et indiquez son chemin d'accès.

Pour spécifier le répertoire dans lequel vous souhaitez le copier, cliquez sur le bouton **[...]** à coté du champ **Destination directory** et sélectionnez le chemin d'accès au fichier original *customer.csv*.

Pour écraser le fichier original par le nouveau fichier trié, cochez la case **Rename** et saisissez *customer.csv* entre les guillemets.

Pour supprimer le fichier temporaire, cochez la case **Remove source File**.



Exécuter le Job

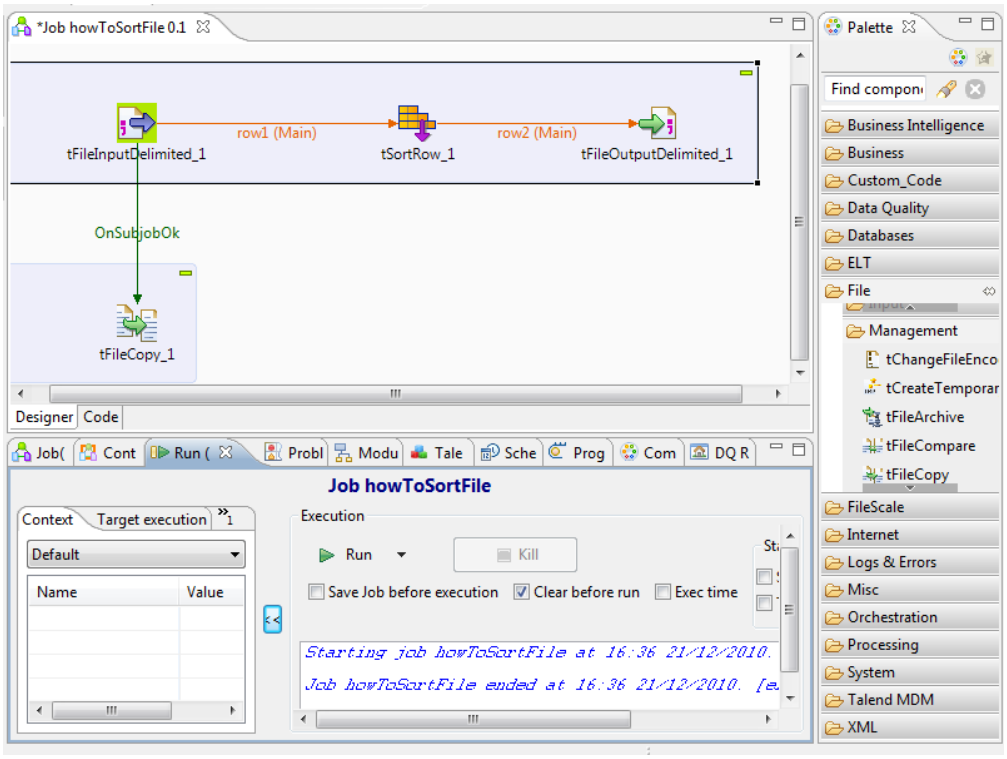
Dans le Job Designer :

Avant d'exécuter votre Job, enregistrez-le via **Ctrl+S**.

Appuyez sur **F6** pour lancer l'exécution.

La vue **Run** s'affiche en bas de **Talend Open Studio** et la console retrace l'exécution du Job.

Exécutez de nouveau ce Job mais en cochant la case **Statistics** de la vue **Run** : cette option permet de mieux comprendre comment sont orchestrés les sous-jobs.



Le Job *howToSortFile* fonctionne !

Il comprend deux sous-jobs permettant de :
- trier des données dans un fichier temporaire,
- remplacer le fichier d'origine par le fichier temporaire.

Il ne nous reste plus qu'à le documenter !

Documenter le Job

Dans le Job Designer :

Pour documenter votre Job, donnez un titre à chacun des sous-jobs.

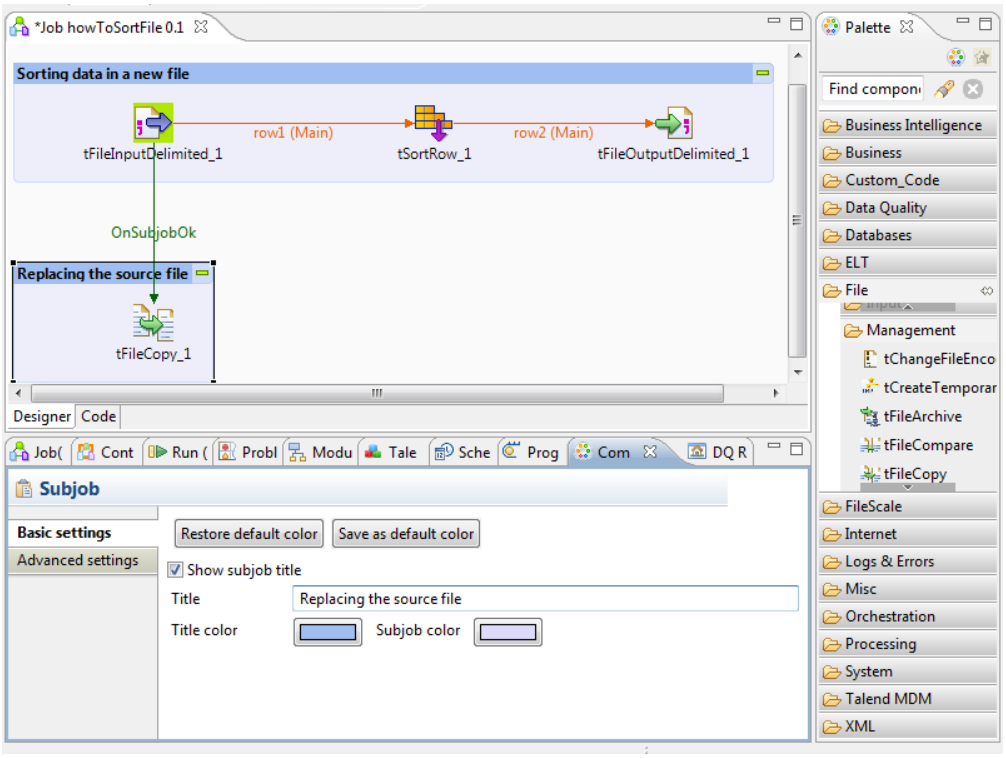
Cliquez sur la zone bleue entourant votre premier sous-job.

Cliquez sur la vue **Component**.

Cochez la case **Show subjob title** et dans le champ **Title** saisissez le titre correspondant : *Sorting data in a new File* (Trier les données dans un nouveau fichier, en français).

De la même manière, donnez le titre *Replacing the source File* (Remplacer le fichier d'origine, en français) à votre deuxième sous-job.

Enregistrez de nouveau votre Job.



B - Créer une métadonnée de connexion à un fichier délimité : Paramétrer des schémas de fichier délimité dans le Repository.

Ce tutoriel montre comment utiliser l'assistant "Delimited File" lorsque vous devez traiter des formats de fichiers complexes. Vous pouvez créer des schémas spécifiques correspondant à chacun de vos besoins.

Par exemple, vous pouvez imaginer un schéma contenant une "adresse personnelle" et un autre contenant une "adresse de livraison" mais correspondant tous les deux au même fichier.

Prérequis :

Pour suivre ce tutoriel, vous avez besoin d'extraire et d'installer les fichiers customer.csv et state.txt du dossier exampleFile.zip

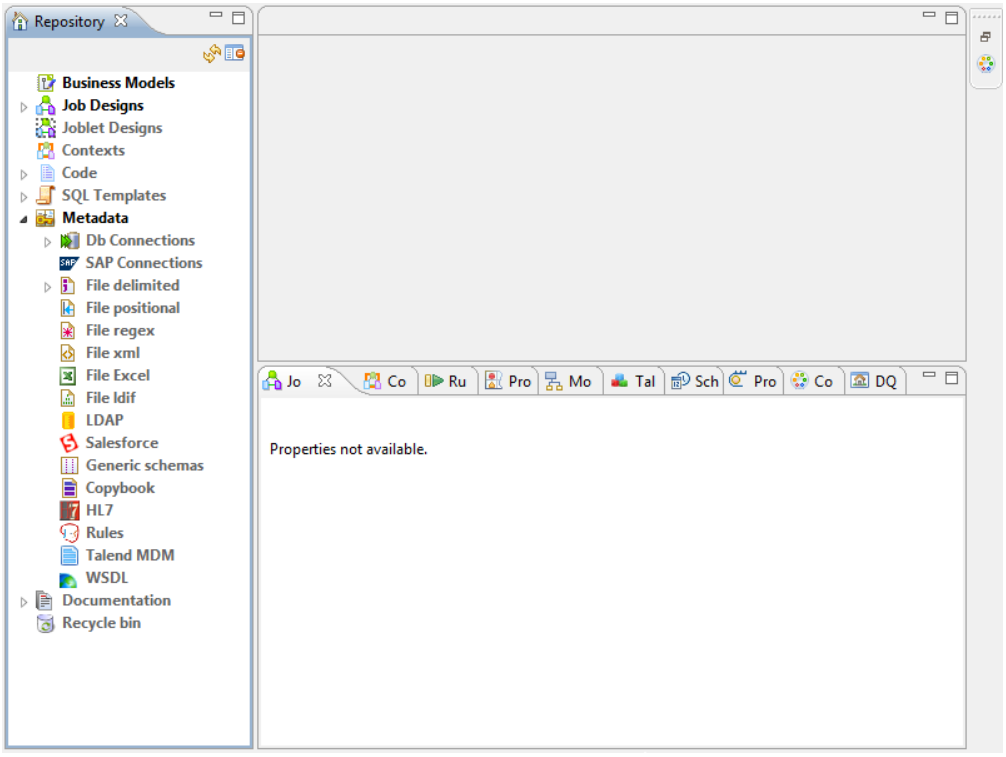
Créer la métadonnée *customers*

Dans le Repository sur la gauche de Talend Open Studio :

Développez le nœud **Metadata**.

Cliquez-droit sur **File delimited**.

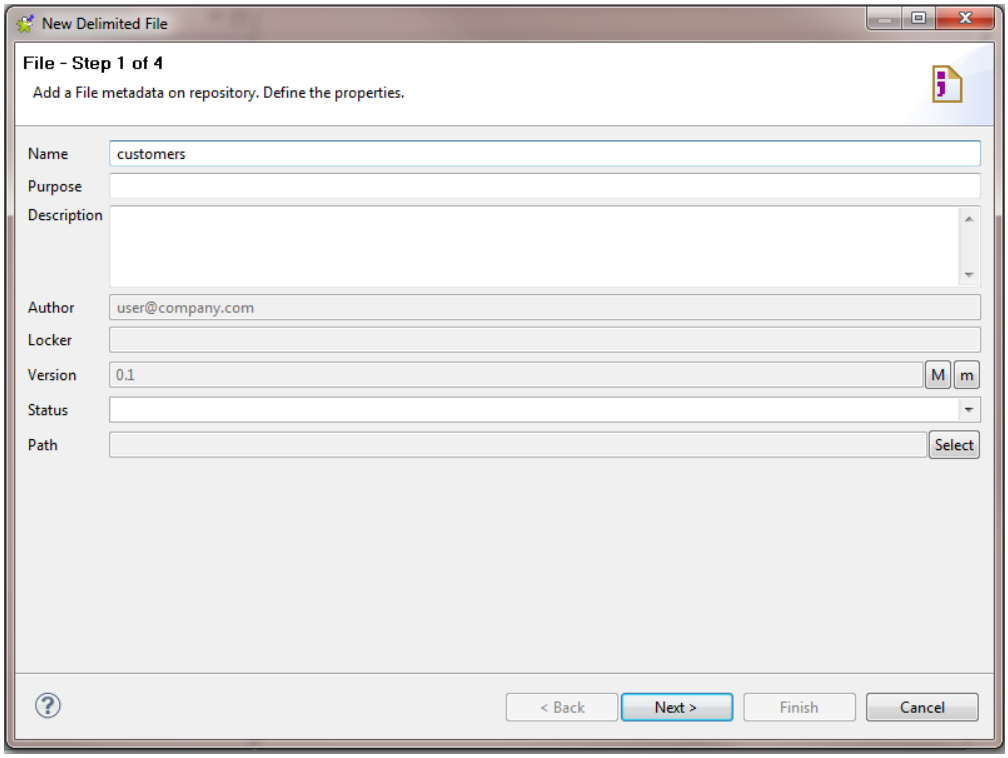
Dans le menu contextuel, cliquez sur **Create file delimited**. L'assistant **New Delimited**



Dans l'assistant New Delimited File :

Dans le champ **Name**, saisissez le nom de la métadonnée : *customer*s.

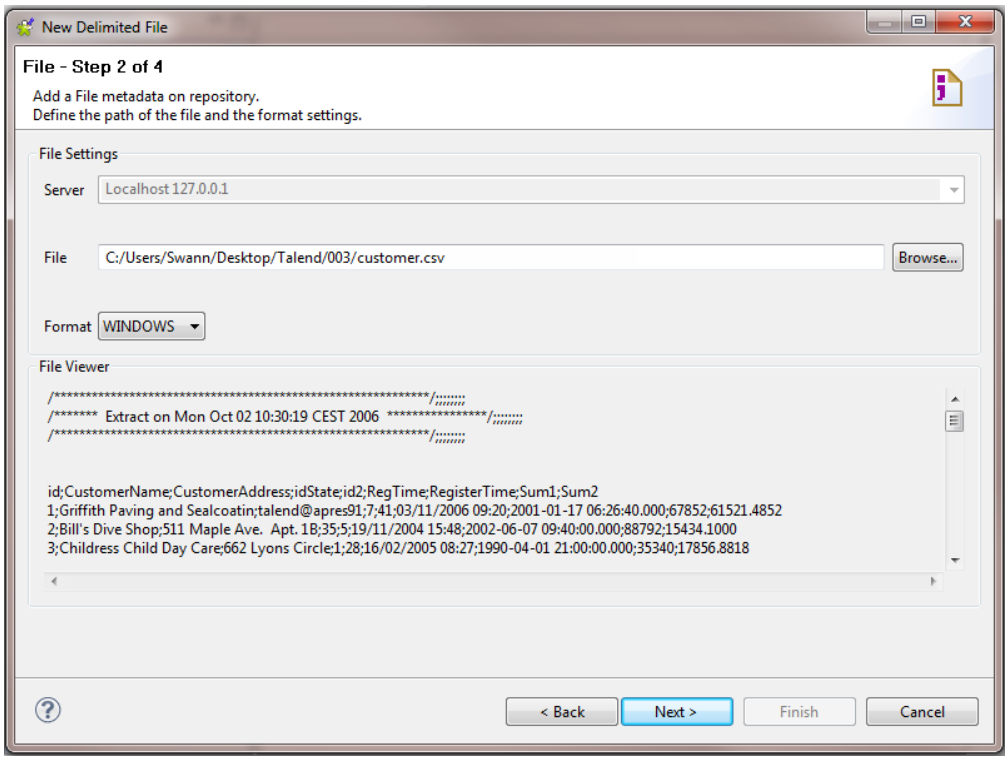
Cliquez sur **Next** pour continuer.



Cliquez sur le bouton **Browse...** et sélectionnez le fichier *customer.csv* dans l'assistant qui s'ouvre.

Dans la liste **Format**, sélectionnez le système d'exploitation de votre ordinateur.

Cliquez sur **Next**.



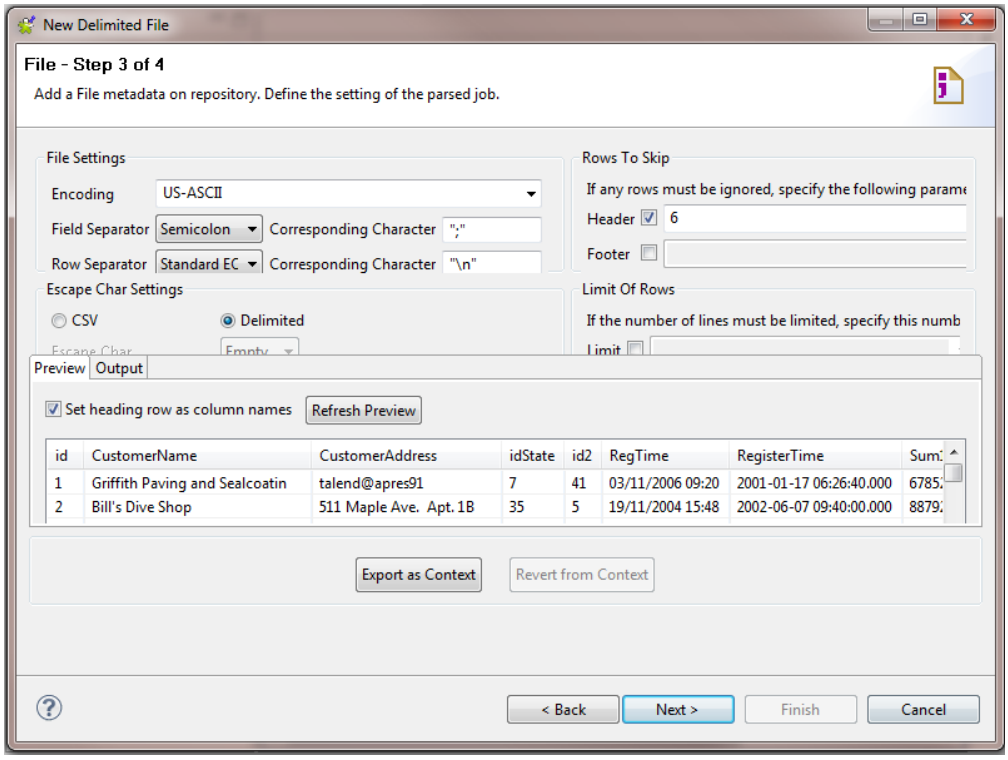
Dans la zone **Rows To Skip**, cochez la case **Header** et saisissez le nombre de lignes d'en-tête à ignorer : *5*.

Dans la zone **Preview** en bas de l'assistant, cochez la case **Set heading row as column names** pour récupérer les noms des colonnes du fichier.

Cliquez sur le bouton **Refresh Preview** pour rafraîchir l'aperçu de la structure et des données du fichier.

Cliquez sur **Next**.

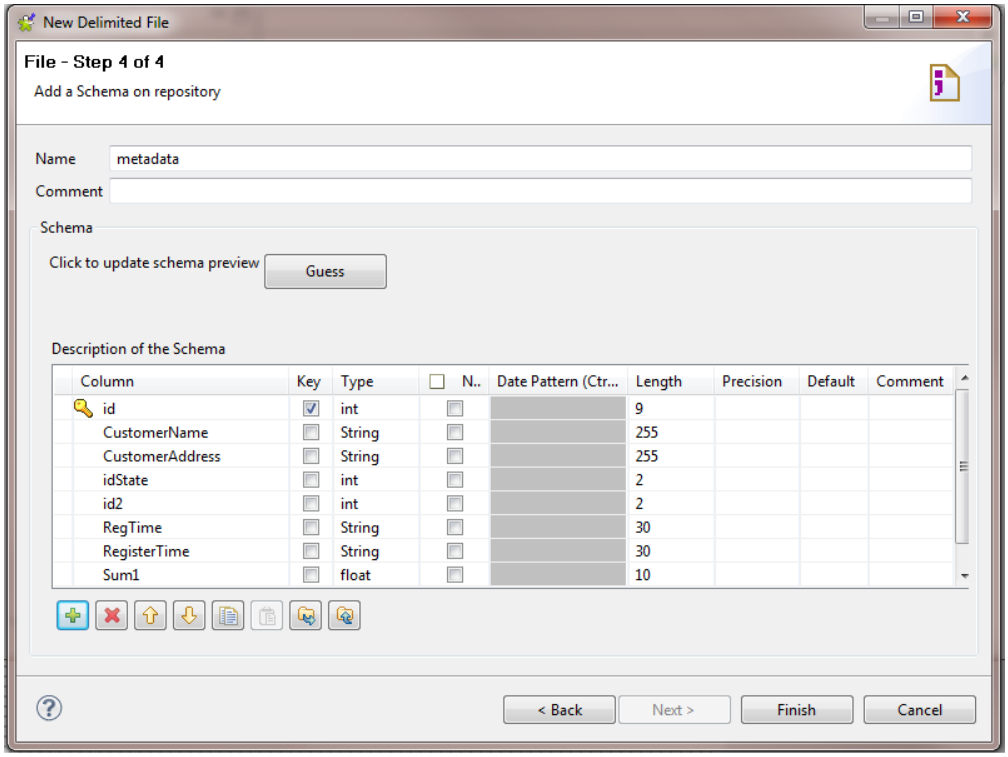
Lorsque vous cochez la case **Set heading row as column names**, le nombre que vous avez saisi dans le champ **Header** est incrémenté.



Dans le tableau **Description of the schema**, paramétrez la clé, le type et la longueur des colonnes comme indiqué sur la capture d'écran.

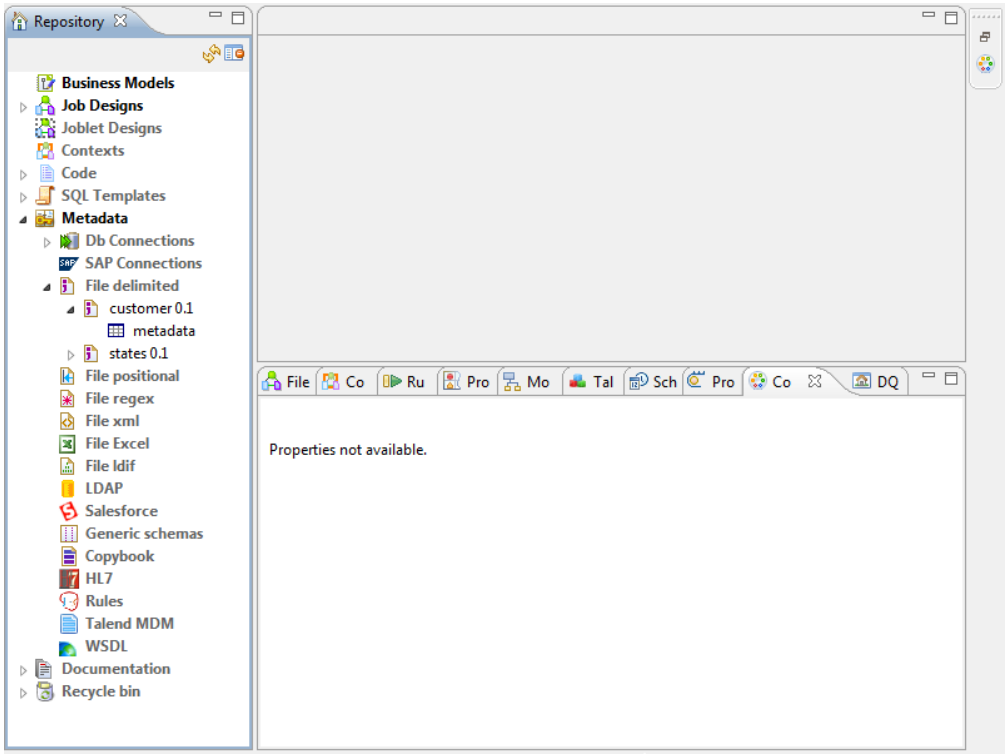
Les colonnes *Sum1* et *Sum2* sont paramétrées de la même manière.

Cliquez sur **Finish** pour fermer l'assistant.



Dans le Repository situé à gauche :

La métadonnée *customers* est affichée sous le nœud **Metadata** > **File delimited**.



La métadonnée *customers* est créée.

Maintenant, créez la métadonnée *states*.

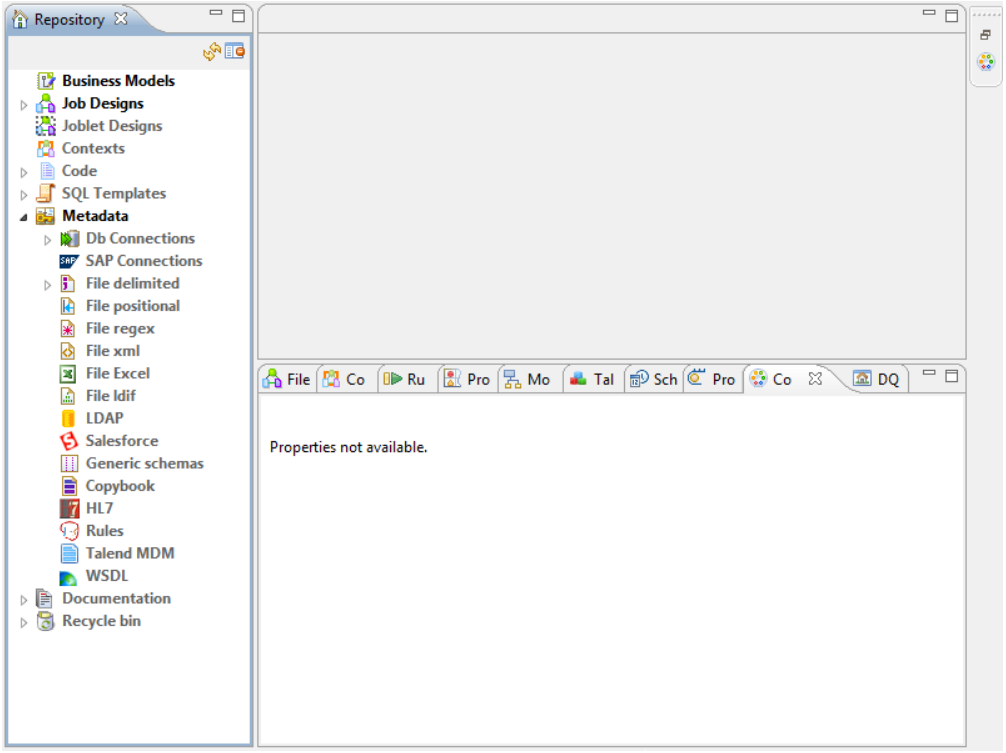
Créer la métadonnée *states*

Dans le Repository sur la gauche de Talend Open Studio :

Développez le nœud **Metadata**.

Cliquez-droit sur **File delimited**.

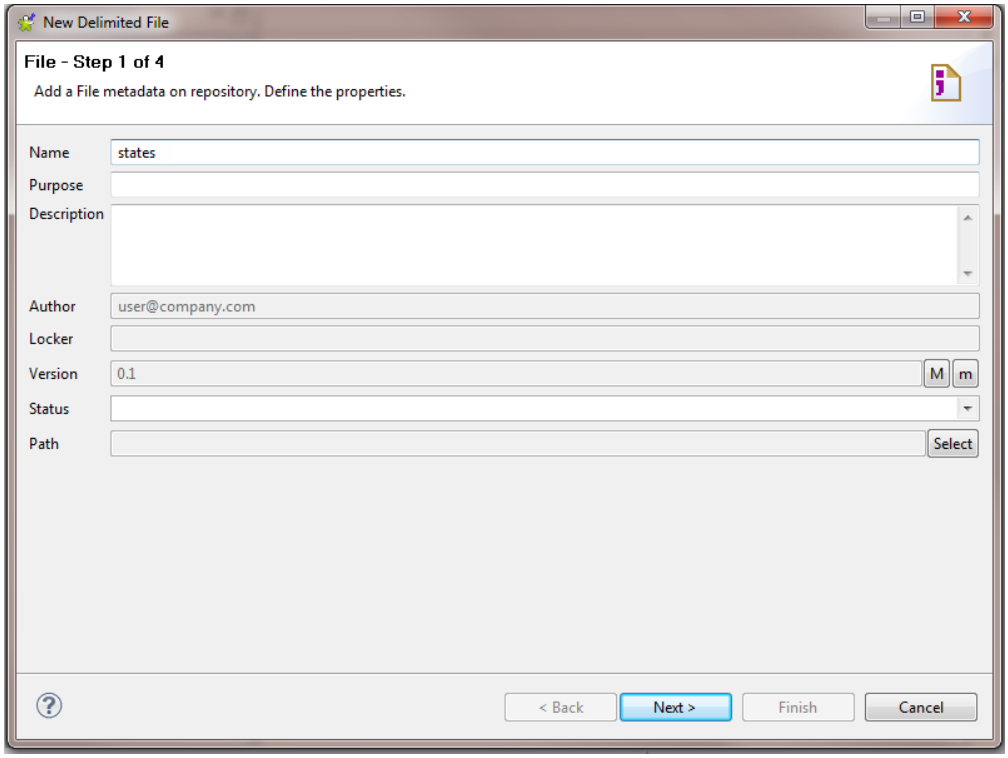
Dans le menu contextuel, cliquez sur **Create file delimited**. L'assistant **New Delimited File** s'ouvre.



Dans l'assistant New Delimited File :

Dans le champ **Name**, saisissez le nom de la métadonnée : *states*.

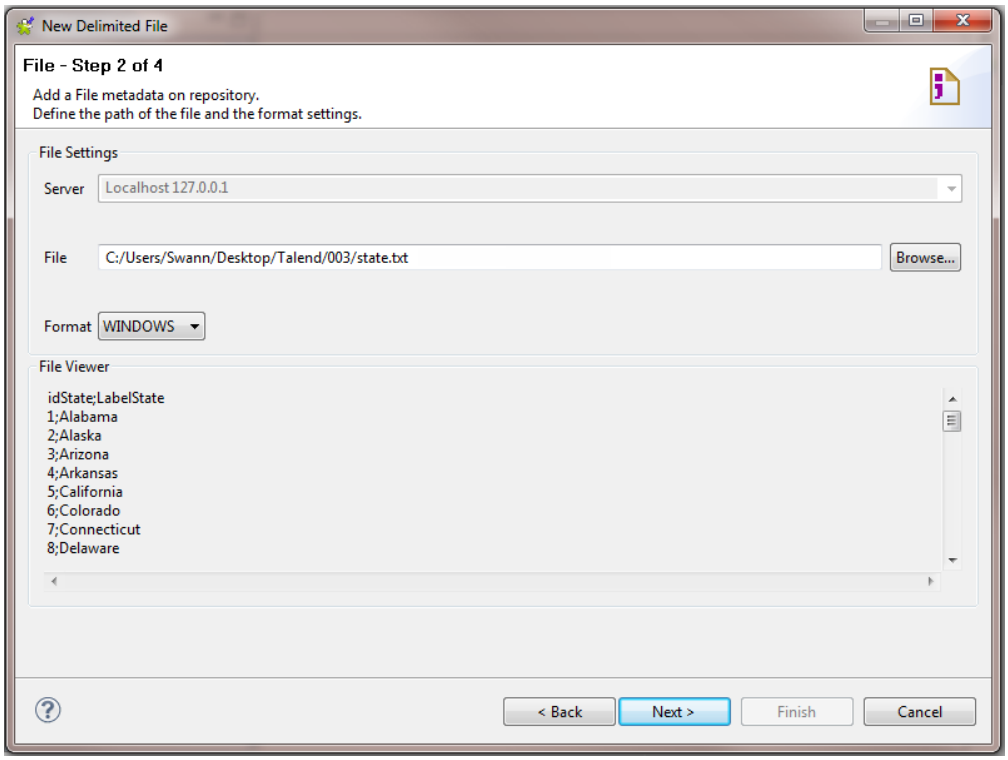
Cliquez sur **Next** pour continuer.



Cliquez sur le bouton **Browse...** et sélectionnez le fichier *state.txt* dans l'assistant qui s'ouvre.

Dans la liste **Format**, sélectionnez le système d'exploitation de votre ordinateur.

Cliquez sur **Next**.



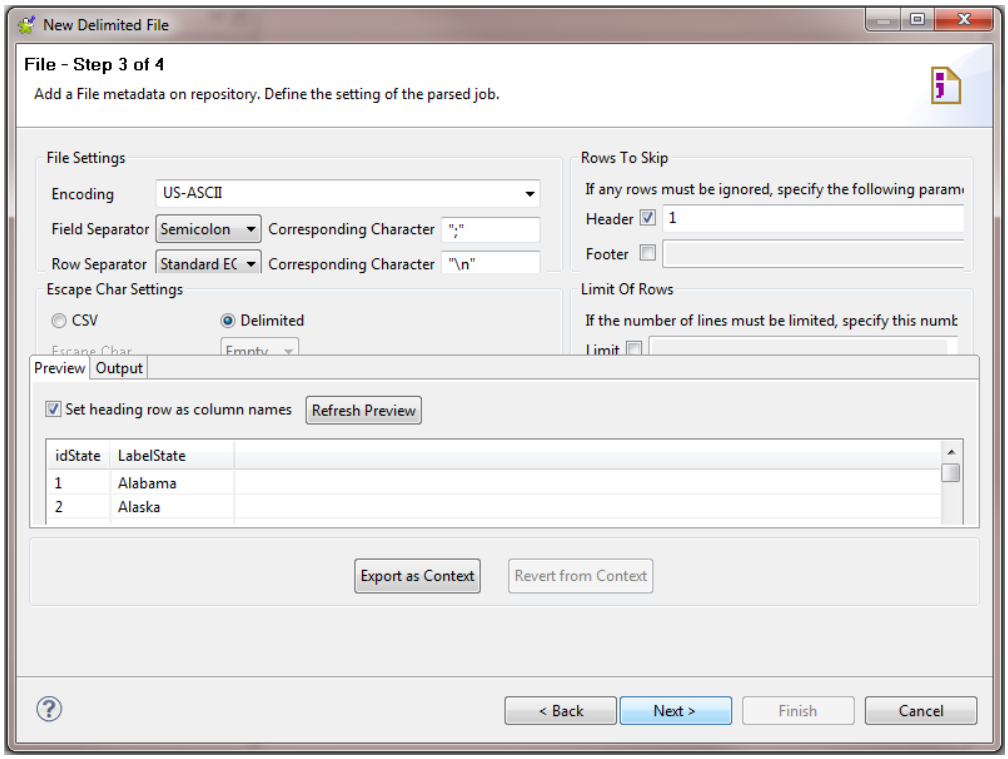
Dans la zone **Preview** en bas de l'assistant, cochez la case **Set heading row as column names** pour récupérer les no

des colonnes du fichier.

Cliquez sur le bouton **Refresh Preview** pour rafraîchir l'aperçu de la structure et des données du fichier.

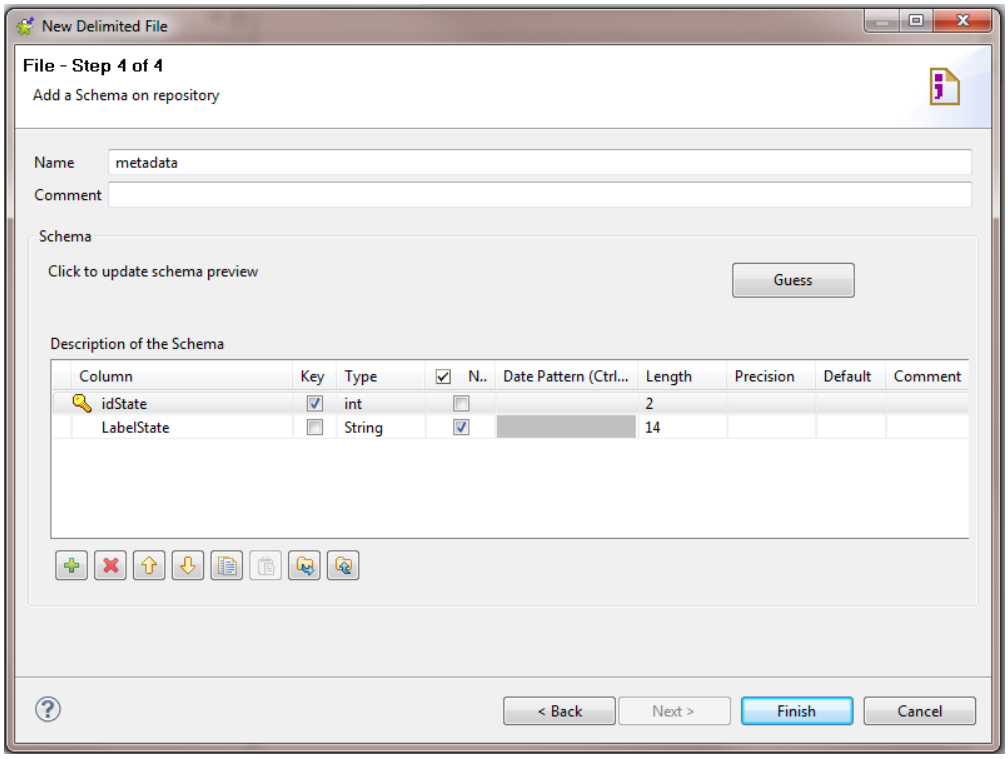
Cliquez sur **Next**.

Lorsque vous cochez la case **Set heading row as column names**, le champ **Header** est incrémenté.



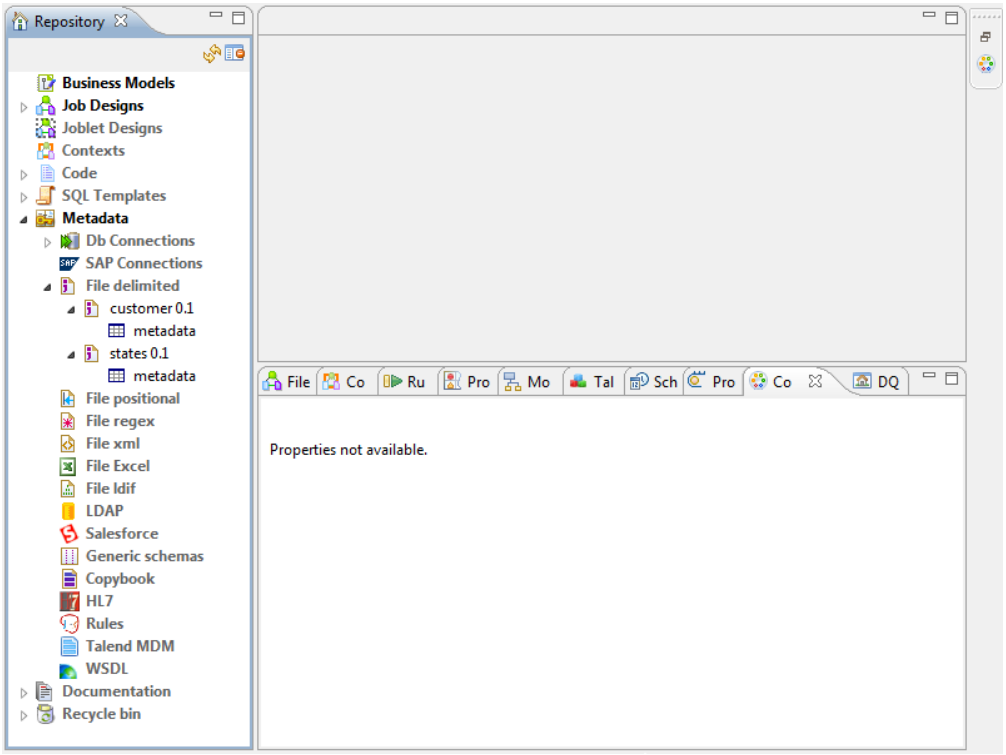
Dans le tableau **Description of the schema**, paramétrez la clé, le type et la longueur des colonnes comme indiqué sur la capture d'écran.

Cliquez sur **Finish** pour fermer l'assistant



Dans le Repository situé à gauche :

La métadonnée *states* s'affiche sous le nœud **Metadata** > **File delimited**



Les deux métadonnées ont été créées. Vous pouvez maintenant les utiliser et réutiliser dans vos Jobs.

C - Créer une jointure dans le Job Designer : Agréger et transformer des données à l'aide du composant tMap.

Dans ce tutoriel, vous utiliserez un fichier customers.csv contenant les informations sur des clients américains et un fichier states.txt contenant la liste de tous les Etats américains avec leur identifiant.

Vous paramétrez le composant tMap pour agréger les données des deux fichiers d'entrée et alimenter le fichier de sortie.

Prérequis :

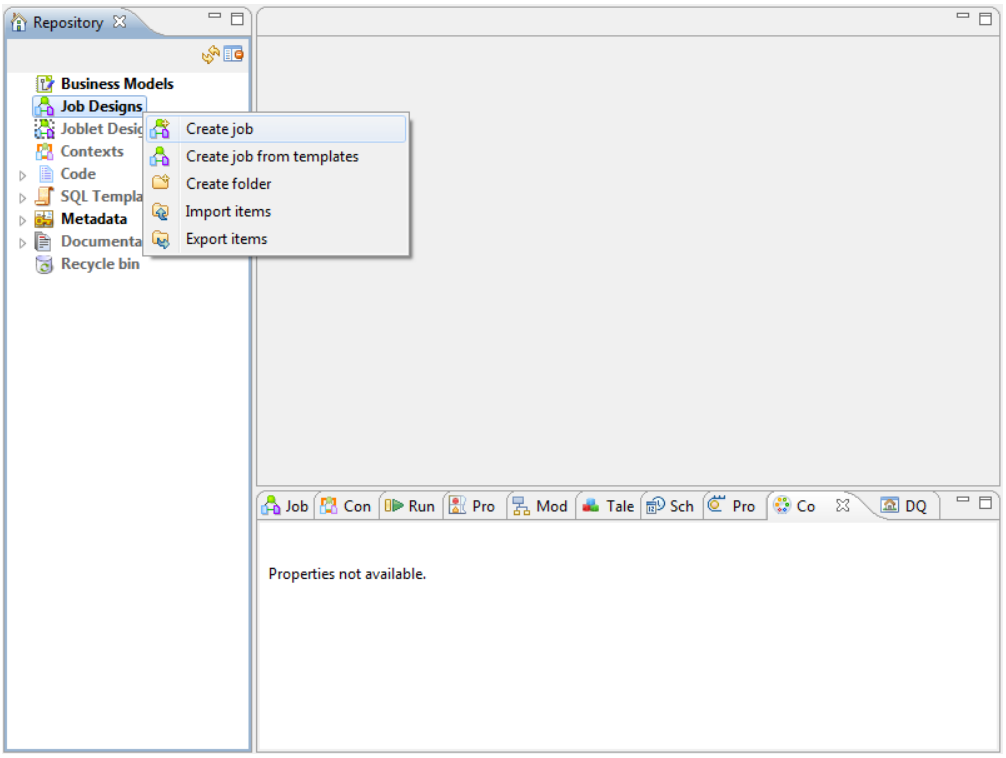
Pour suivre ce tutoriel, vous avez besoin d'extraire et d'installer les fichiers customer.csv et state.txt du dossier exampleFile.

Créer le Job Design

Dans le Repository situé à gauche de Talend Open Studio :

Pour créer un Job, cliquez-droit sur **Job Designs**.

Dans le menu contextuel, cliquez sur **Create Job** pour ouvrir l'assistant **New Job**.



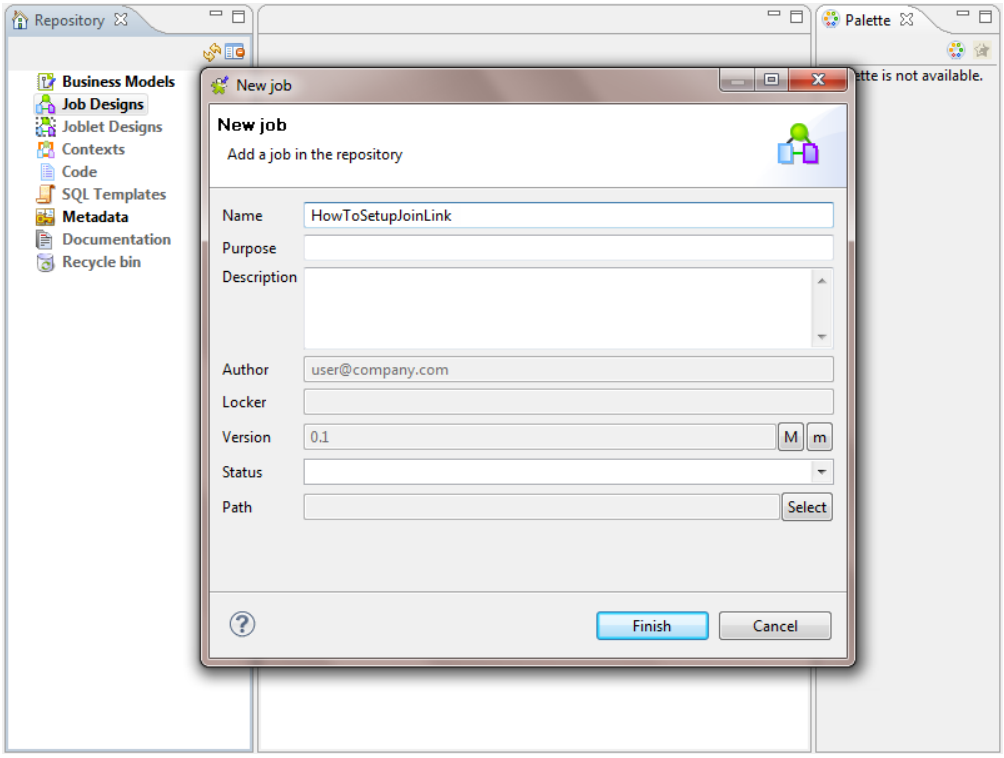
Dans l'assistant New Job

Dans le champ **Name**, saisissez le nom du Job : *HowtoSetupJoinLink*.

Cliquez sur **Finish** pour fermer l'assistant et créer votre Job.

Le Job Designer présente alors un Job vierge.

Le champ **Name** ne doit pas contenir d'accents, de caractères spéciaux, d'espaces, ni débuter par un chiffre.



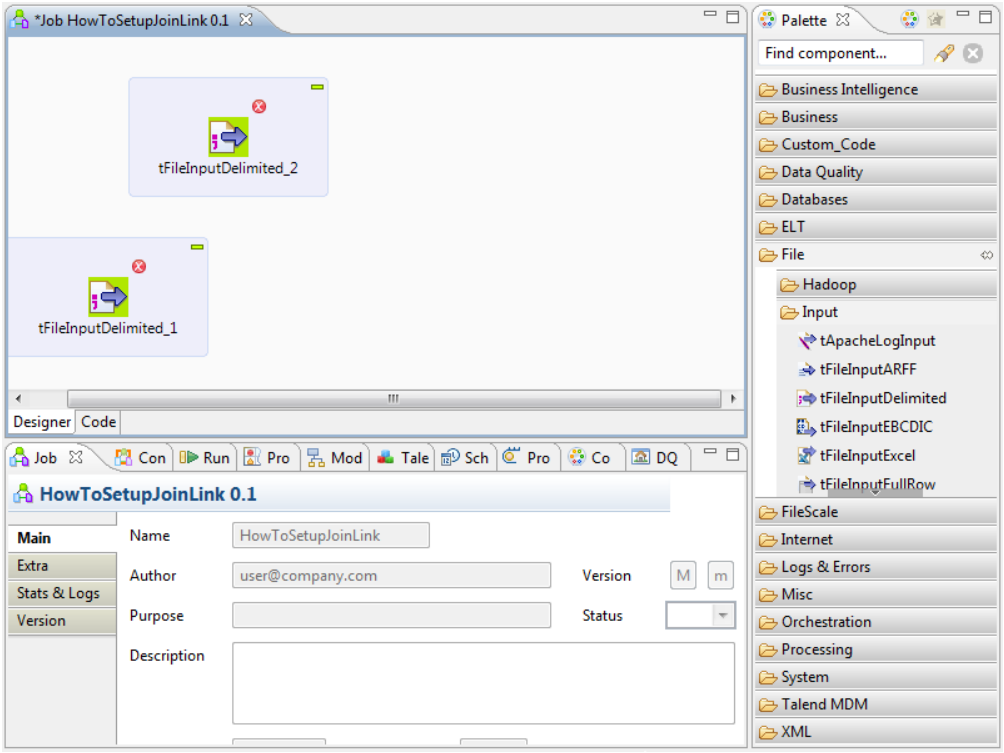
Paramétrer les connecteurs de lecture de fichiers délimités

Dans la Palette située à droite :

Pour ajouter le composant d'entrée, cliquez sur la famille **File** et sur la sous-famille **Input**.

Cliquez sur le composant **tFileInputDelimited** et déposez-le dans le Job Designer.

De la même manière, ajoutez un deuxième composant **tFileInputDelimited**.



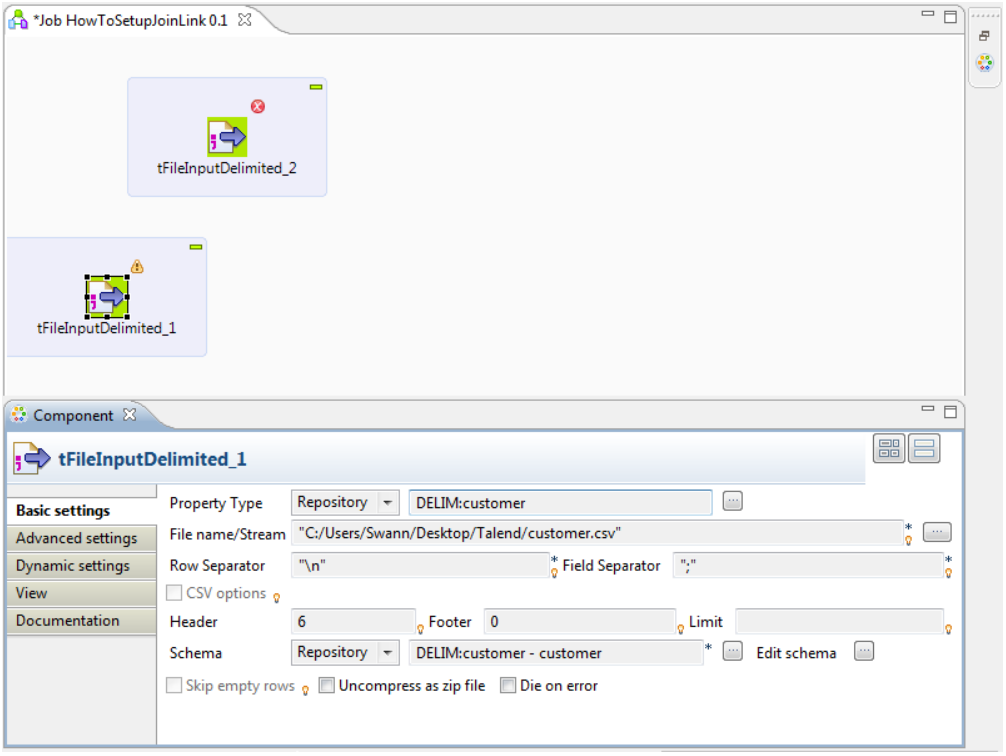
Dans le Job Designer :

Pour paramétrer les propriétés du **tFileInputDelimited_1**, double-cliquez sur le composant. La vue **Component** correspondante apparaît alors en bas de l'écran.

Dans la vue Component :

Pour spécifier les propriétés du composant, sélectionnez **Repository** dans la liste **Property Type** puis cliquez sur le bouton [...] situé à coté du champ **Edit schema** pour vérifier le schéma du fichier.

L'assistant **Edit parameter using repository** s'ouvre.

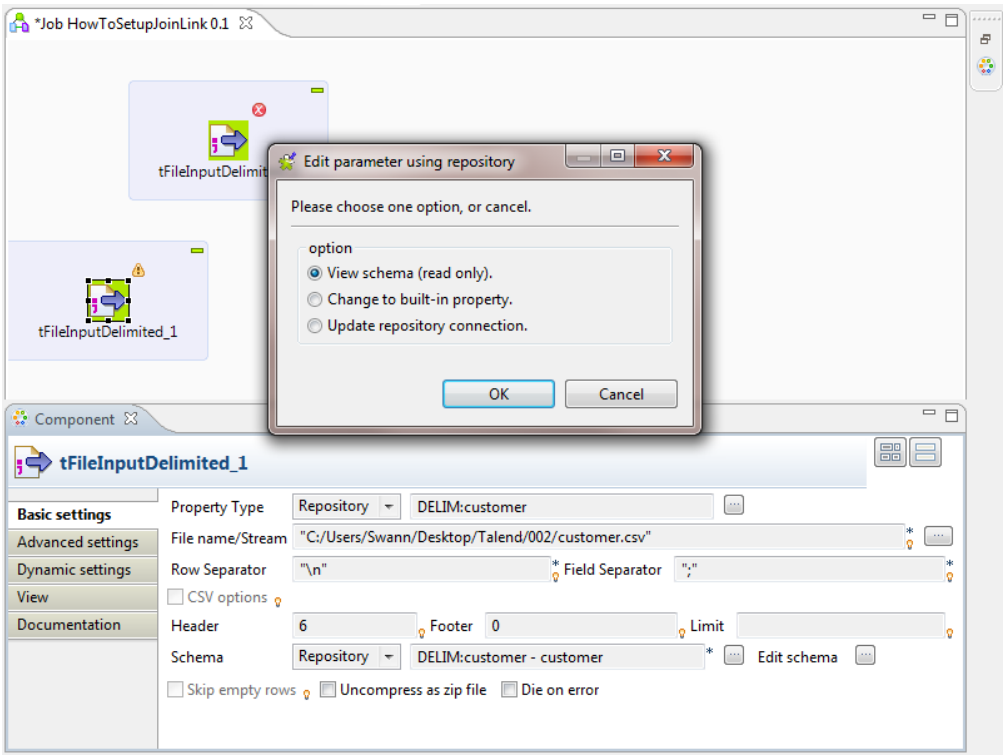


Dans l'assistant Edit parameter using repository :

Sélectionnez **View schema (read only)** dans la liste **option** puisque vous ne souhaitez que consulter le schéma.

Cliquez sur **OK**.

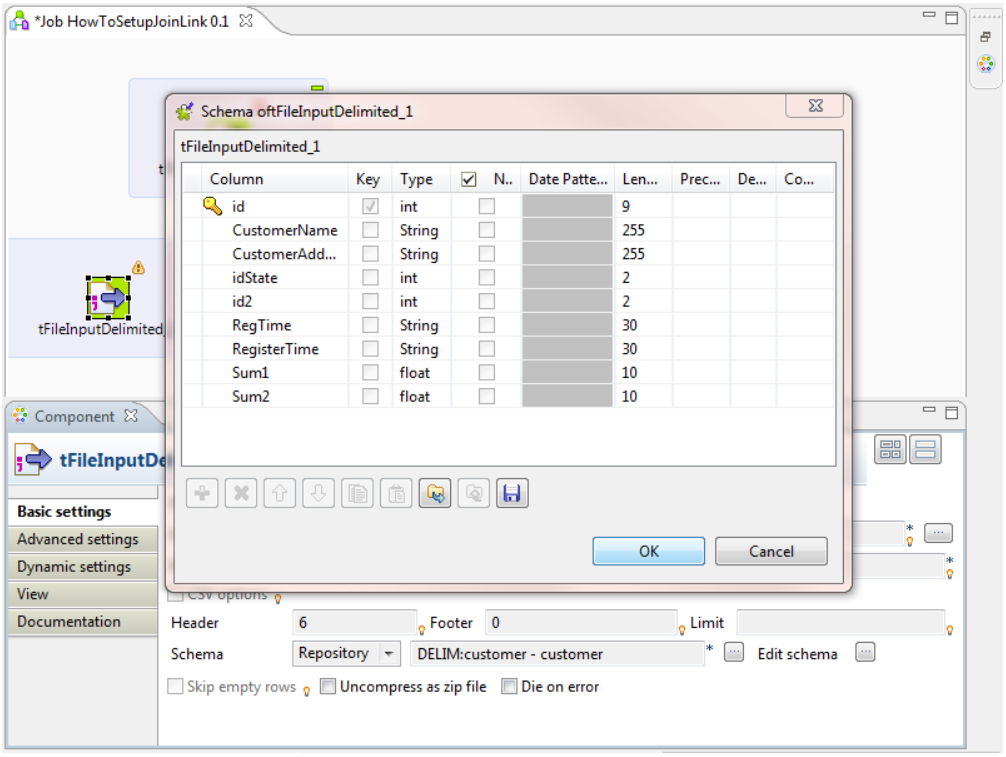
L'assistant **Schema of tFileInputDelimited_1** s'ouvre.



Dans l'assistant Schema of tFileInputDelimited_1 :

Le schéma est le même que celui disponible à partir de la vue **Repository** sous le nœud **Metadata** > **File delimited**.

Cliquez sur **OK**.



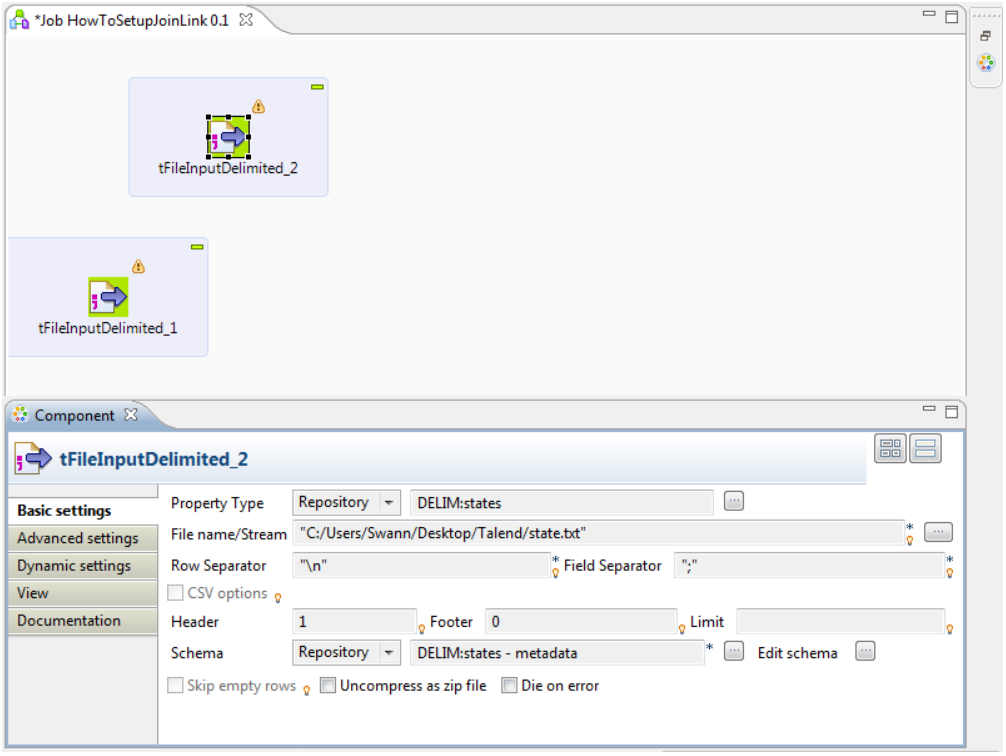
Dans le Job Designer :

Pour paramétrer les propriétés du **tFileInputDelimited_2**, double-cliquez sur le composant. La vue **Component** correspondante apparaît alors en bas de l'écran.

Dans la vue Component :

Pour spécifier les propriétés du composant, sélectionnez **Repository** dans la liste **Property Type** puis cliquez sur le bouton [...].

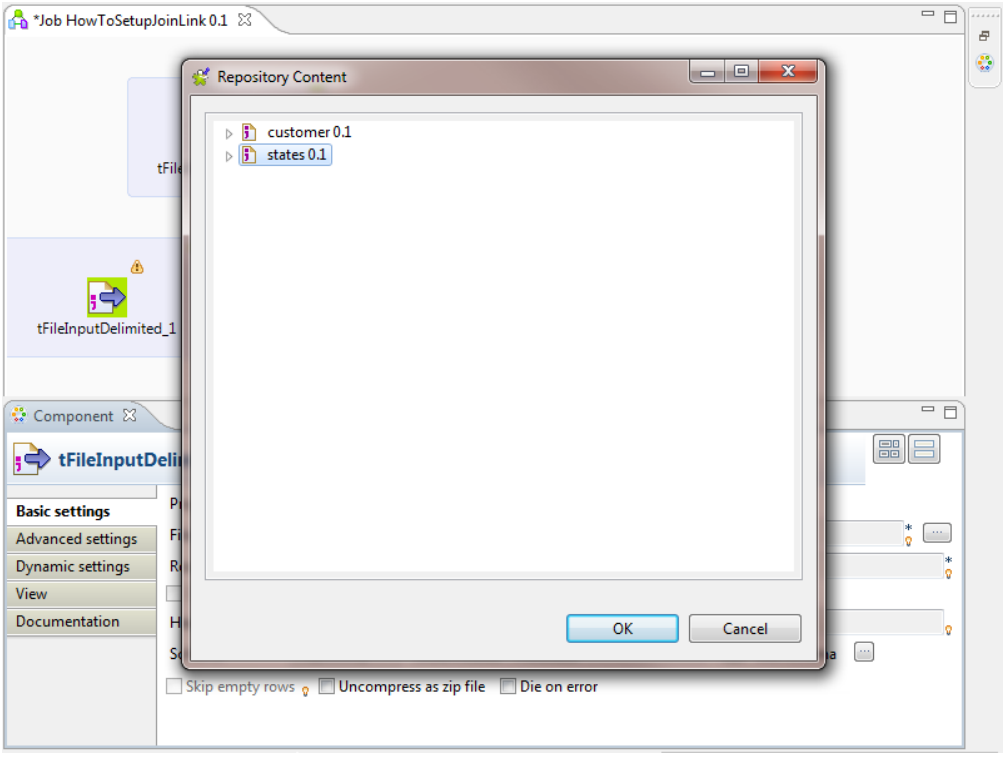
L'assistant **Repository Content** s'ouvre.



Dans l'assistant Repository Content :

Sélectionnez la métadonnée *states* pour que les propriétés du composant **tFileInputDelimited_2** soient automatiquement renseignées avec les propriétés du fichier *states.txt*.

Cliquez sur **OK**.



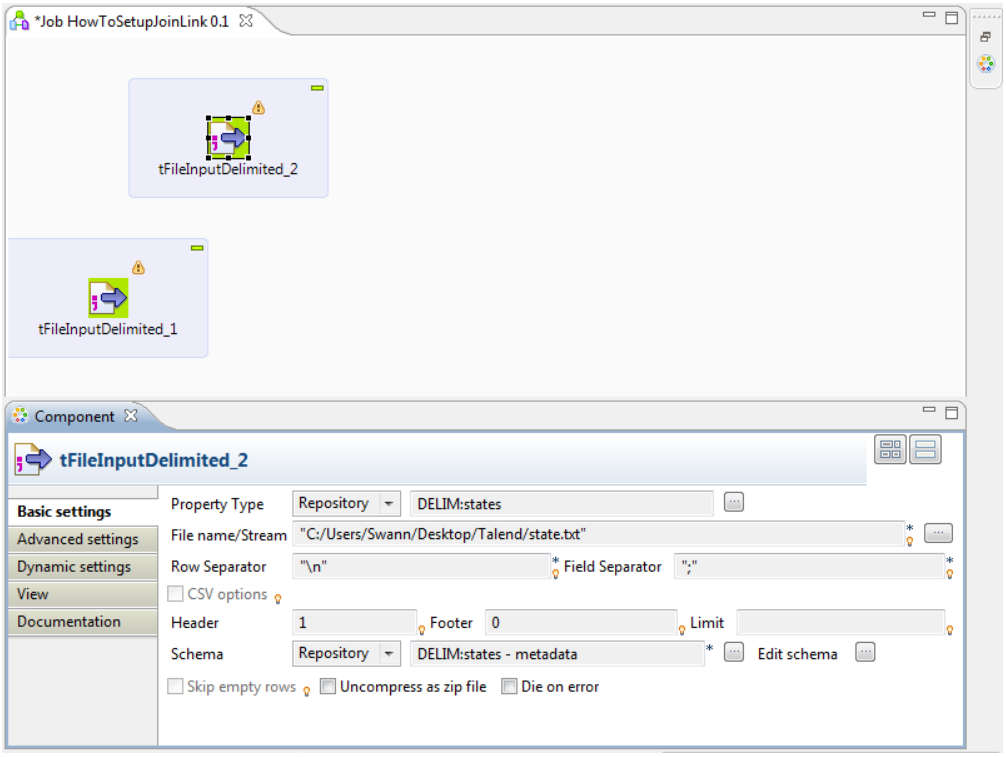
Dans la vue Component :

Cliquez sur le bouton **[...]** situé à coté du champ **Edit schema** pour vérifier le schéma de votre fichier.

L'assistant **Edit parameter using repository** s'ouvre.

Conservez l'option **View schema (read only)** puisque vous ne souhaitez que consulter le schéma et cliquez sur **OK**.

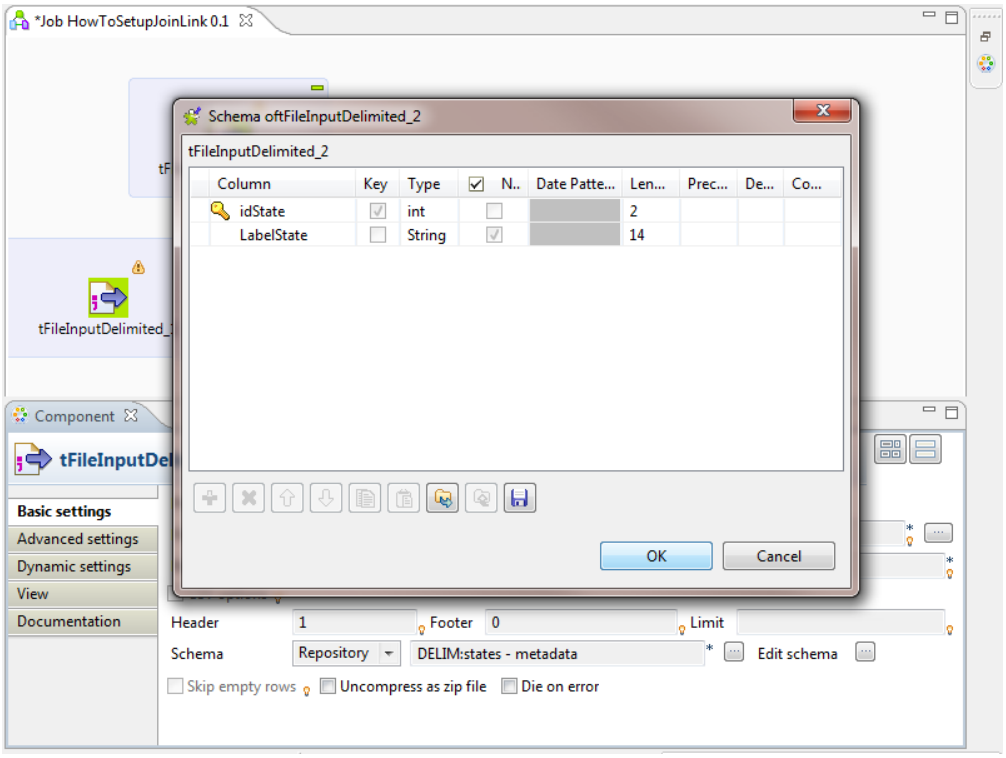
L'assistant **Schema of tFileInputDelimited_2** s'ouvre.



Dans l'assistant Schema of tFileInputDelimited_2 :

Le schéma est le même que celui disponible à partir de la vue **Repository** sous le nœud **Metadata** > **File delimited**.

Cliquez sur **OK**.



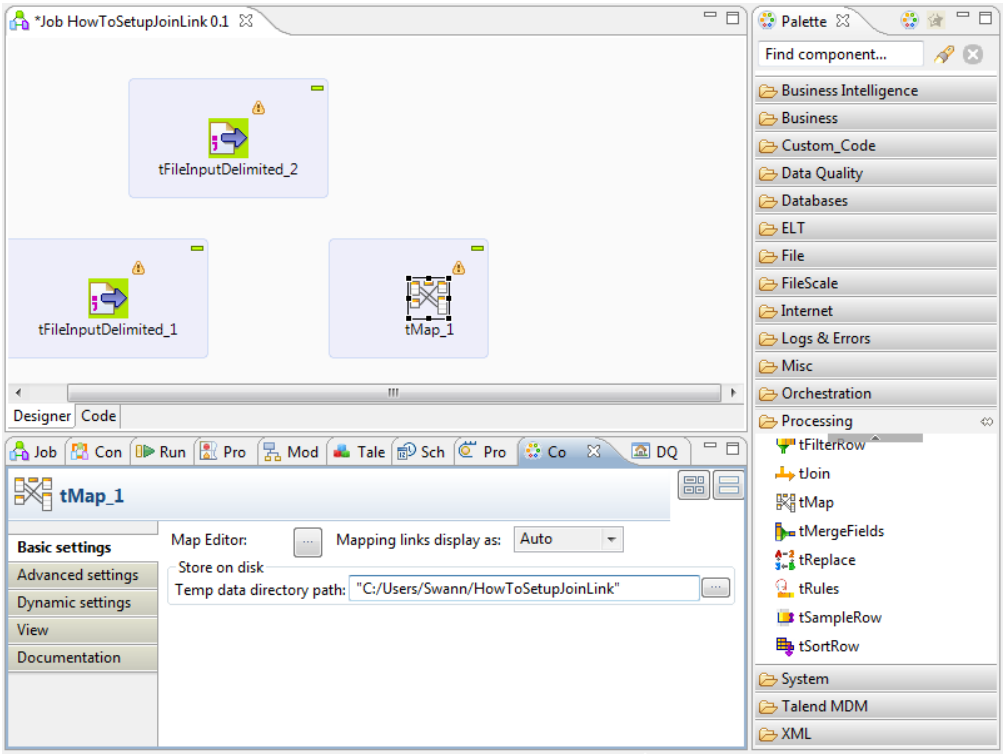
Déposer le composant de transformation et le connecteur d'écriture de fichier délimité dans le Job Designer

---

Dans la Palette située à droite :

Pour ajouter le composant de transformation, cliquez sur la famille **Processing**.

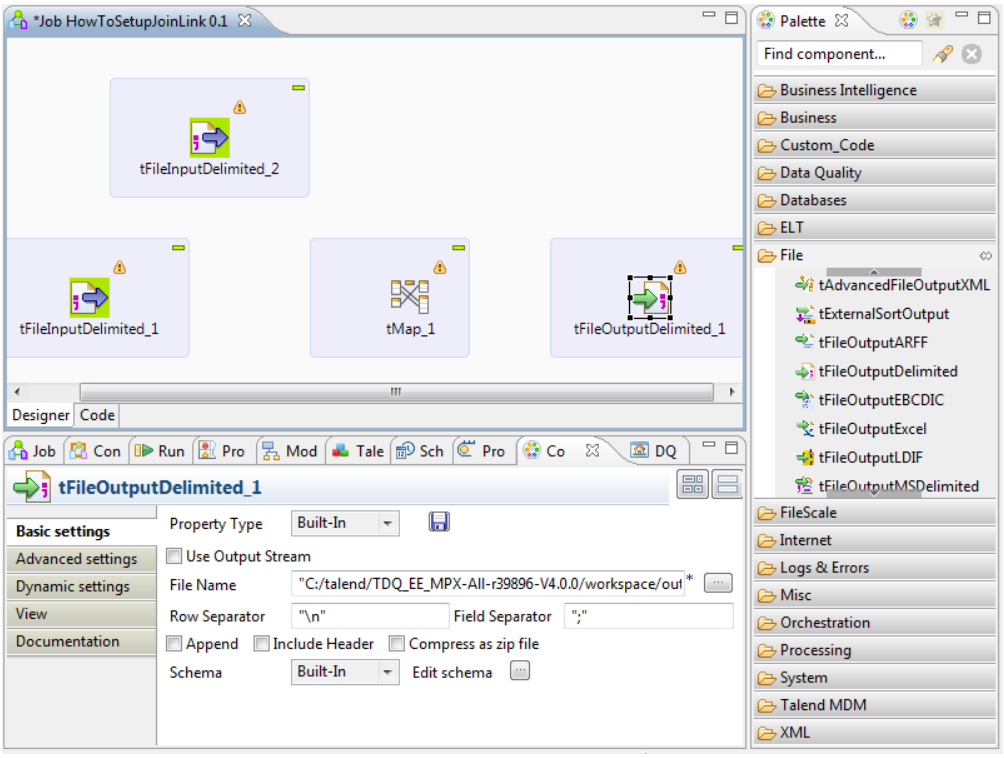
Cliquez sur le composant **tMap** et déposez-le dans le Job Designer.



Dans la Palette située à droite :

Pour ajouter le composant de sortie, cliquez sur la famille **File** et sur la sous-famille **Output**.

Cliquez sur le composant **tFileOutputDelimited** et déposez-le dans le Job Designer.



Relier les composants entre eux

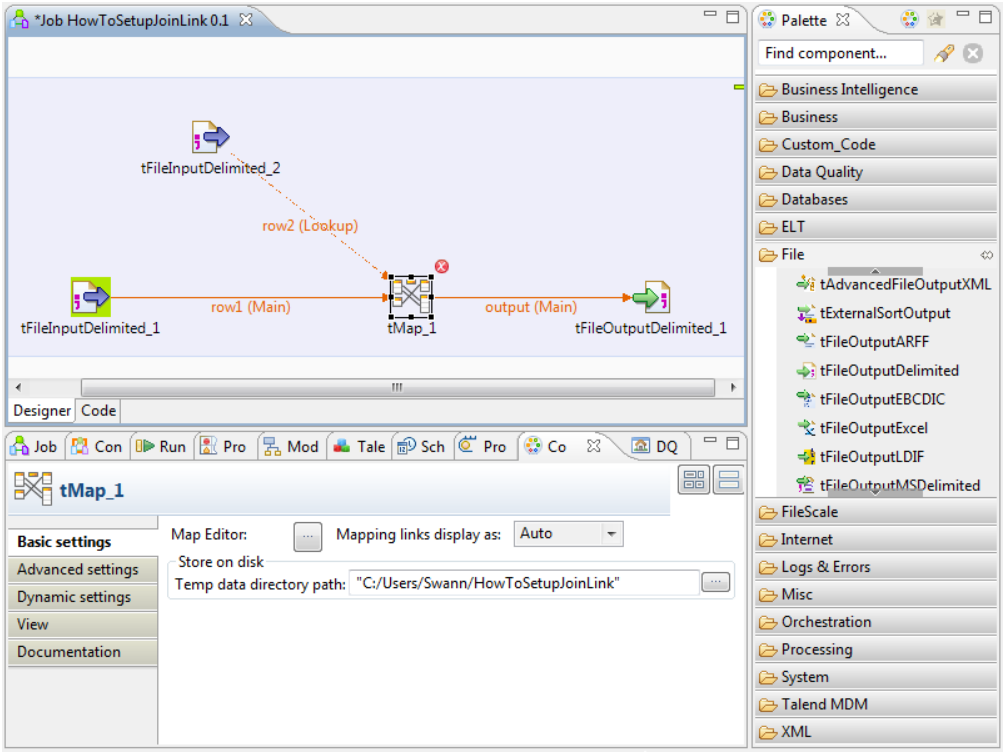
Dans le Job Designer :

Cliquez-droit sur le **tFileInputDelimited_1** et déplacez-vous jusqu'au **tMap**.

De la même manière, créez un lien du **tFileInputDelimited_2** vers le **tMap** puis du **tMap** vers le **tFileOutputDelimited**.

Dans l'assistant **tMap_1 Output**, donnez le nom *output* au lien reliant le **tMap** au **tFileOutputDelimited**.

Vous pouvez aussi créer ce lien en cliquant-droit sur le composant, en sélectionnant **Row** > **Main** dans le menu contextuel et en cliquant sur le composant de sortie.

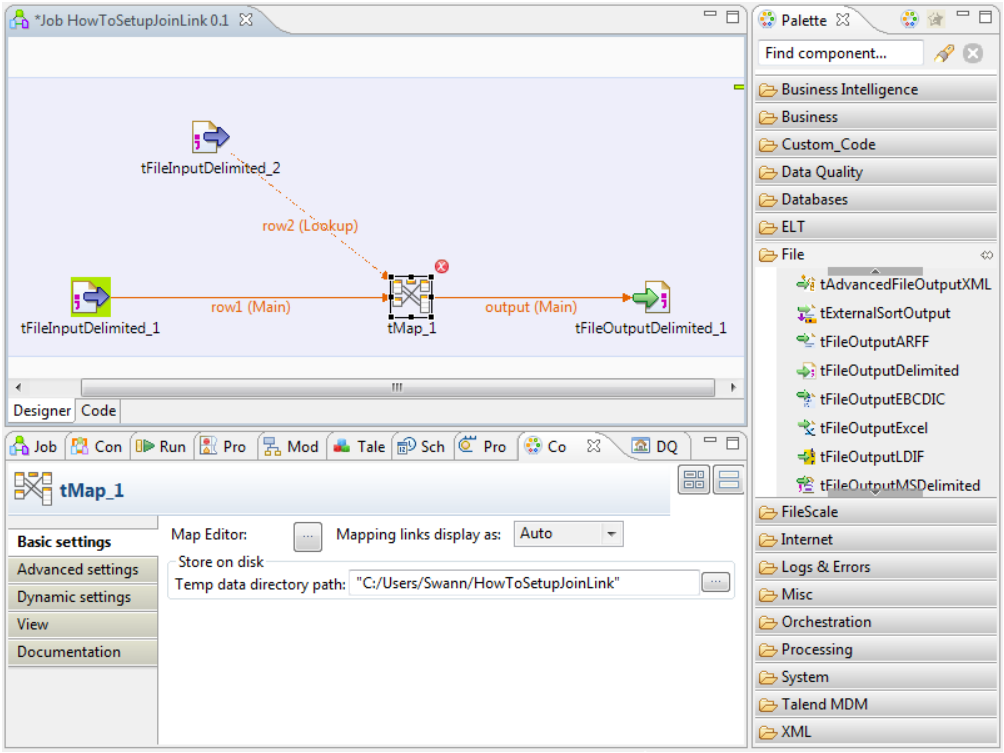


Paramétrer le composant de transformation et le composant d'écriture de fichier délimité

Dans le Job Designer :

Pour paramétrer les propriétés du **tMap**, double-cliquez sur le composant et l'éditeur du tMap s'ouvre.

Vous pouvez aussi sélectionner le **tMap** dans le Job Designer, cliquez sur la vue **Component** et cliquez sur le bouton [...] situé à coté du champ **Map Editor** pour ouvrir l'éditeur du tMap.



Dans l'éditeur du tMap :

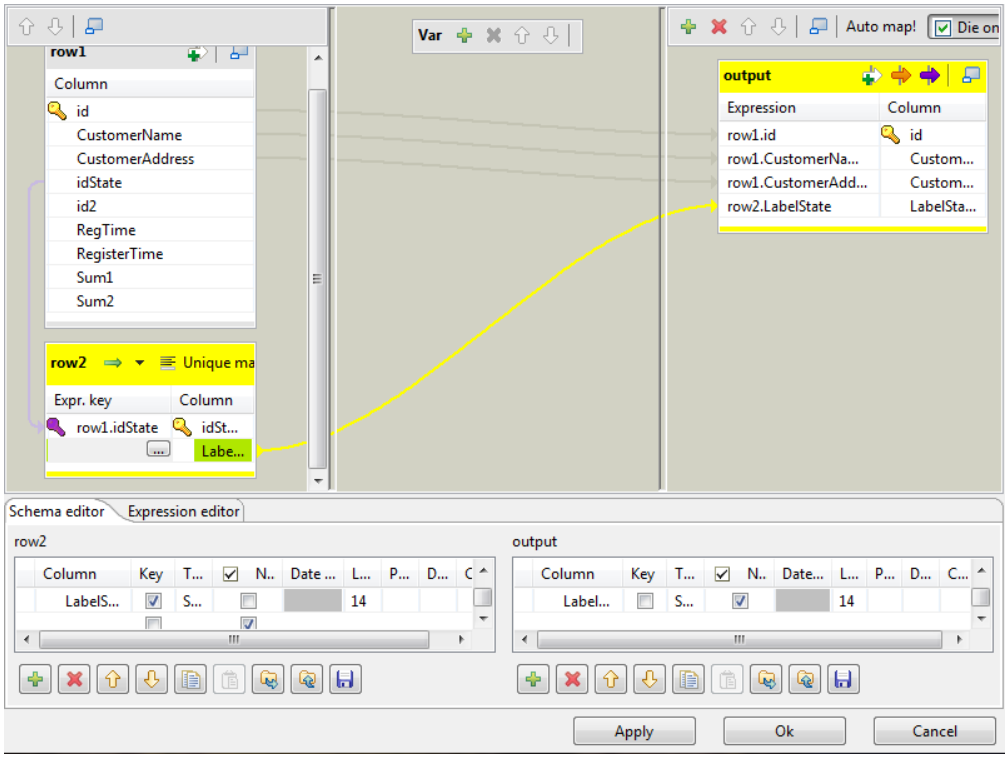
Dans la table **row1**, sélectionnez les colonnes *id*, *CustomerName* et *CustomerAddress* et glissez-les dans la table **output**.

Puis sélectionnez la colonne *idState* et glissez-la dans la colonne *idState* de la table **row2** pour joindre les deux tables.

Dans la table **row2**, sélectionnez la colonne *LabelSate* et glissez-la dans la table **output**.

Cliquez sur **OK**.

La boîte de dialogue **Propagate** s'ouvre. Cliquez sur **Yes** pour propager le schéma défini dans l'éditeur du tMap au composant suivant.



Dans le Job Designer :

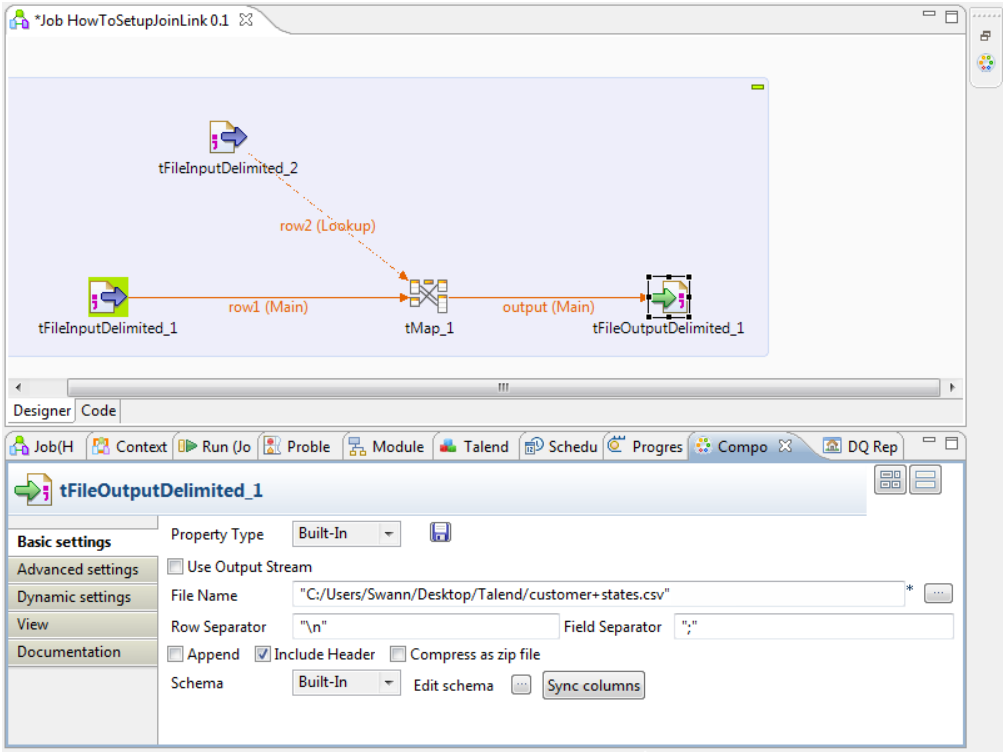
Pour paramétrer les propriétés du **tFileOutputDelimited**, double-cliquez sur ce composant. La vue **Component** correspondante s'affiche alors en bas de l'écran.

Dans la vue Component :

Pour spécifier le chemin du ficher qui sera créé, cliquez sur le bouton [...] à coté du champ **File Name**.

Dans l'assistant, définissez le même répertoire que les fichiers *customer* et *state* et nommez le fichier *customers+states.csv*.

Cochez la case **Include Header** pour récupérer les en-têtes des fichiers d'entrée.



Exécuter le Job

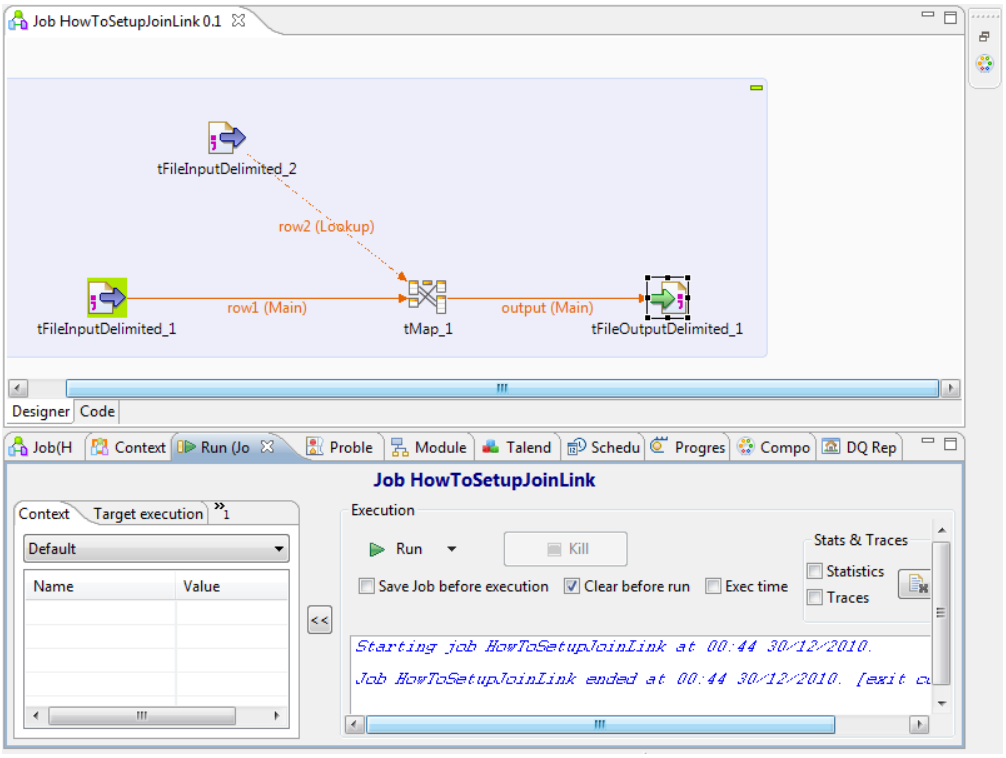
Dans le Job Designer :

Avant d'exécuter votre Job, enregistrez-le via **Ctrl+S**.

Appuyez sur **F6** pour lancer l'exécution du Job.

La vue **Run** s'affiche en bas de **Talend Open Studio** et la console retrace l'exécution du Job.

Exécutez de nouveau ce Job mais en cochant la case **Statistics** de la vue **Run**.



Le Job HowtoSetupJoinLink est presque terminé !

Il permet d'agréger les données des deux fichiers d'entrée et d'alimenter le fichier de sortie.

Il ne reste plus qu'à le documenter !

Documenter le Job

Dans le Job Designer :

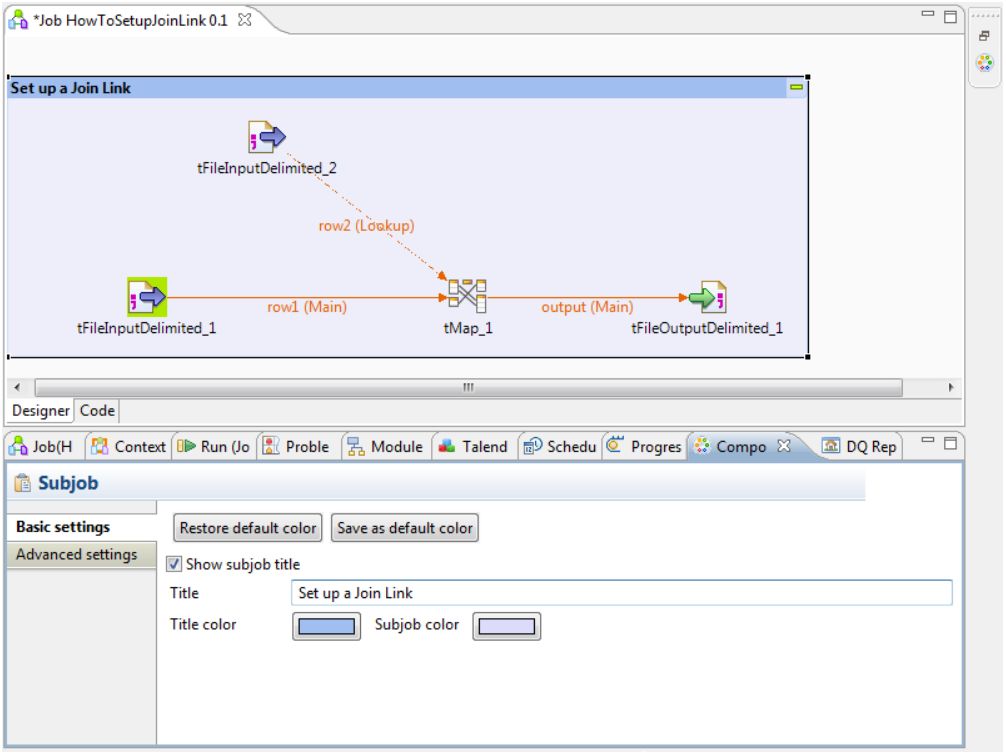
Pour documenter votre Job, donnez-lui un titre.

Pour cela, cliquez sur la zone bleue entourant votre Job.

Cliquez sur la vue **Component**.

Pour lui ajouter un titre, cochez la case **Show subjob title** et dans le champ **Title**, saisissez le titre correspondant : *Set up a Join link* (Créer une jointure, en français).

Enregistrez de nouveau votre Job.



Ce tutoriel est maintenant fini.

Le job fonctionne et est documenté.