Talend Open Studio for Data Integration

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.

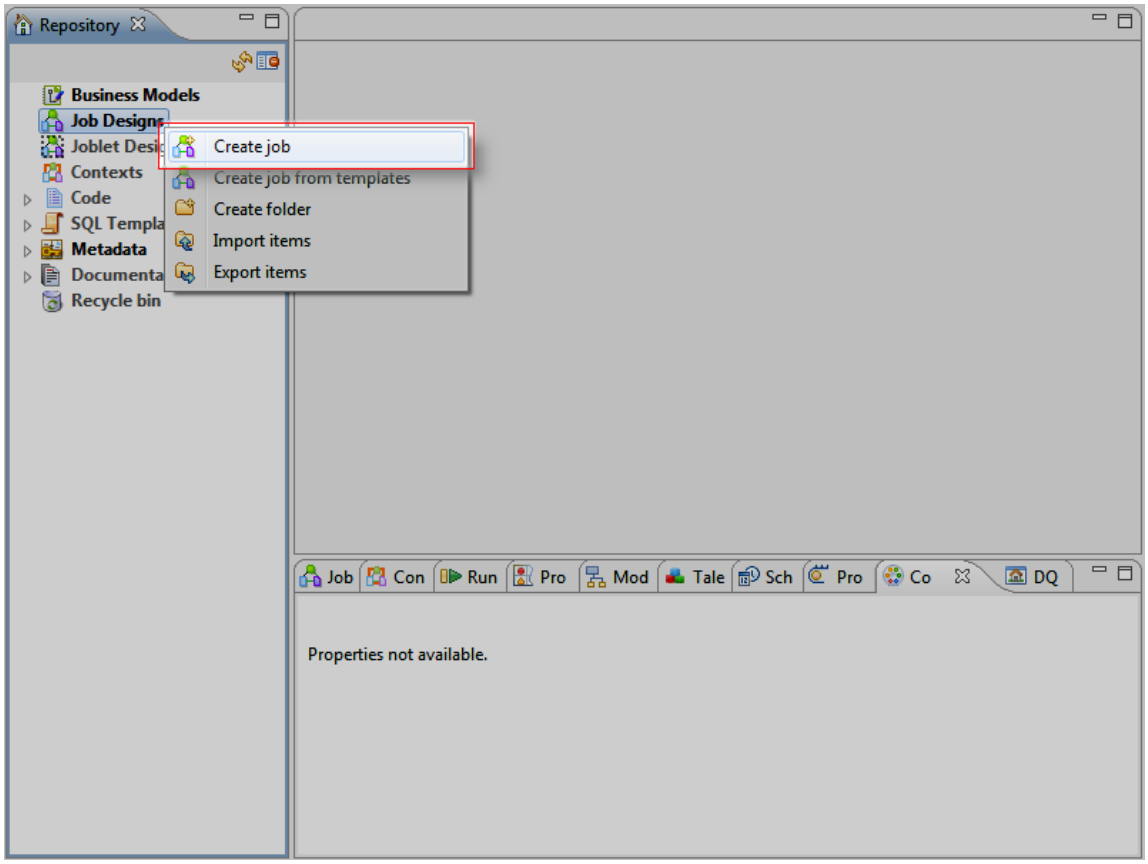
Prérequis :
Pour suivre ce tutoriel, vous avez besoin d'extraire et d'importer le fichier customer.csv.

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**.



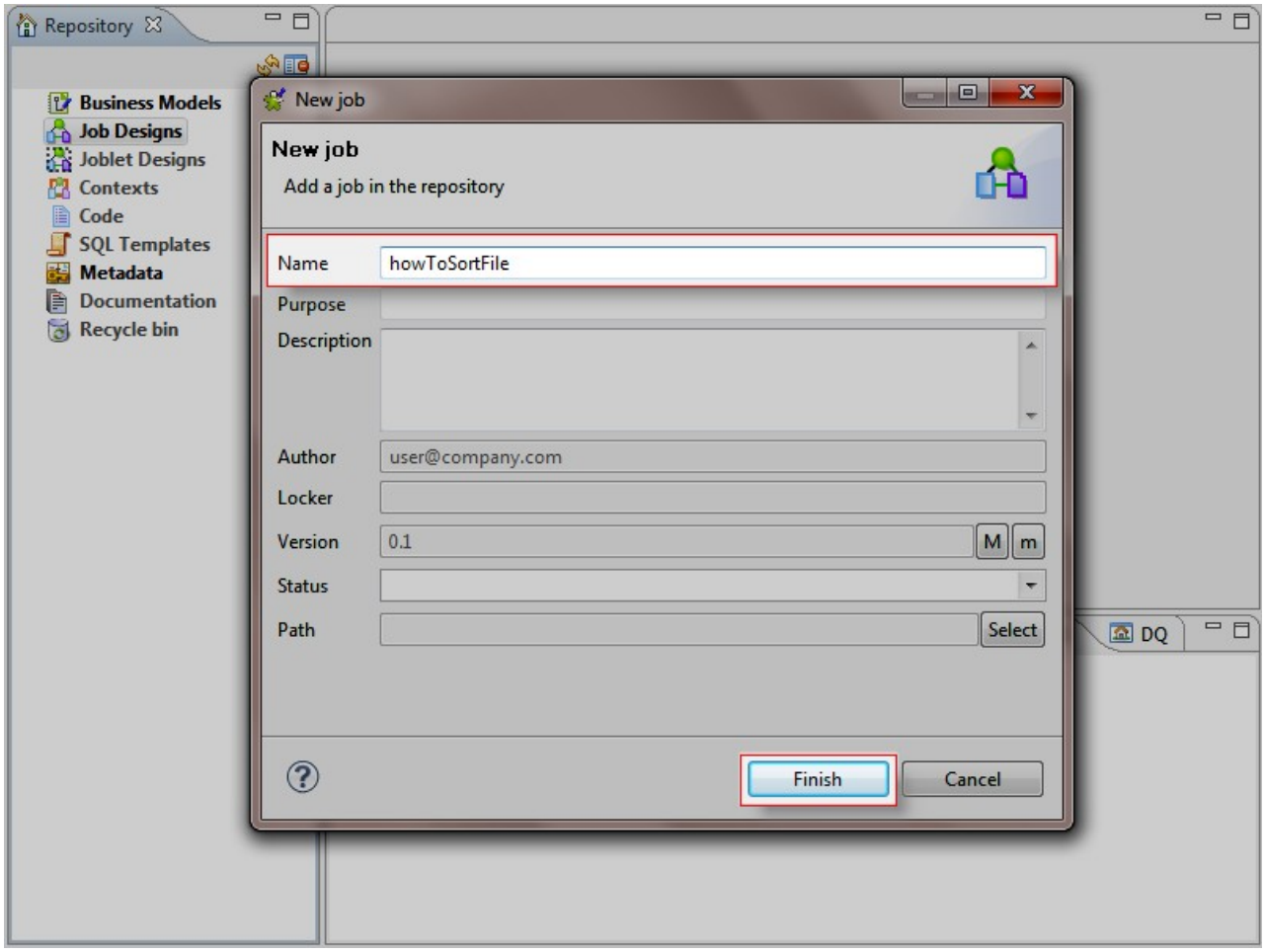
Dans l'assistant New Job :

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.

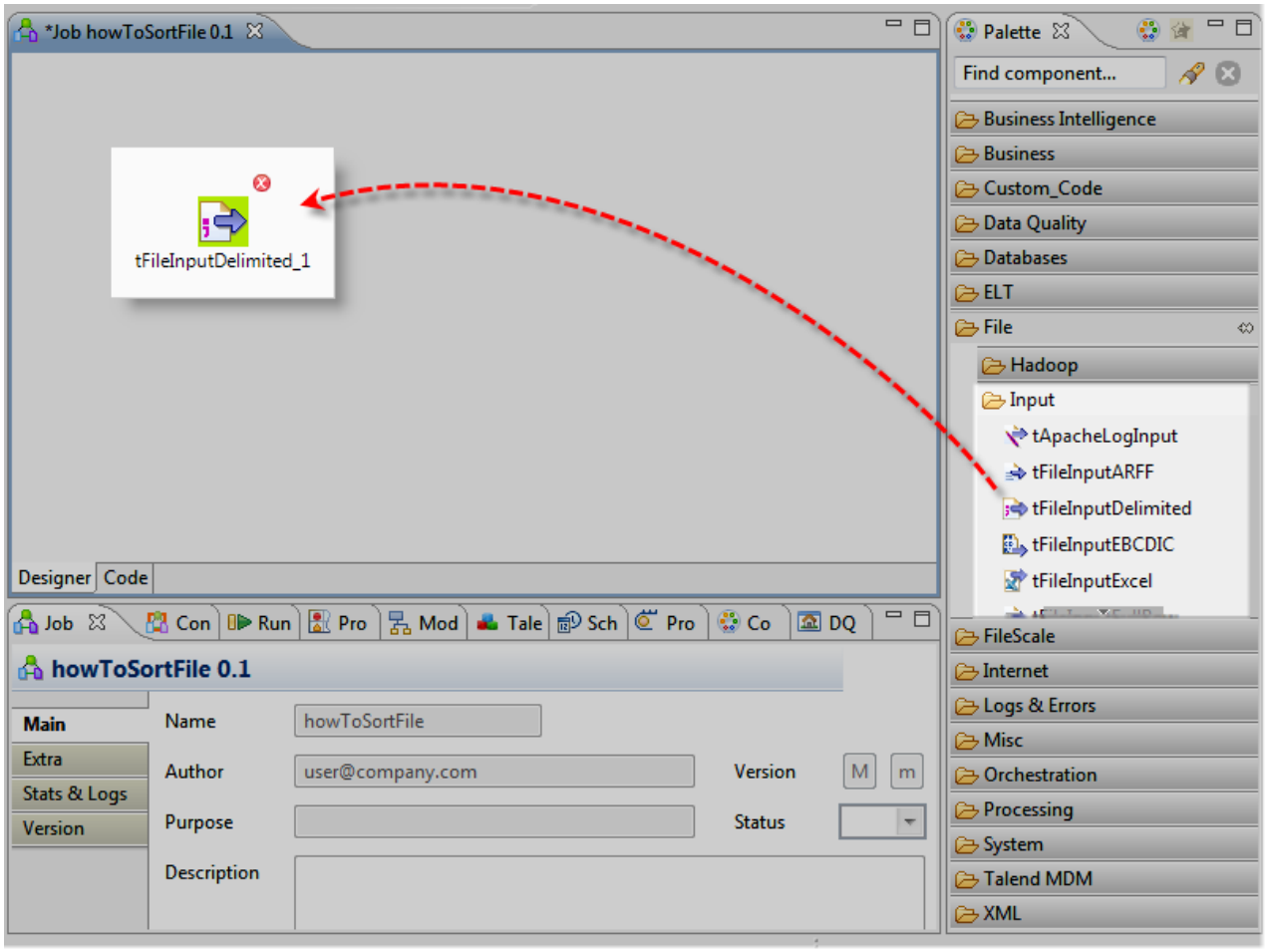


Paramétrer le connecteur de lecture de fichier délimité

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.



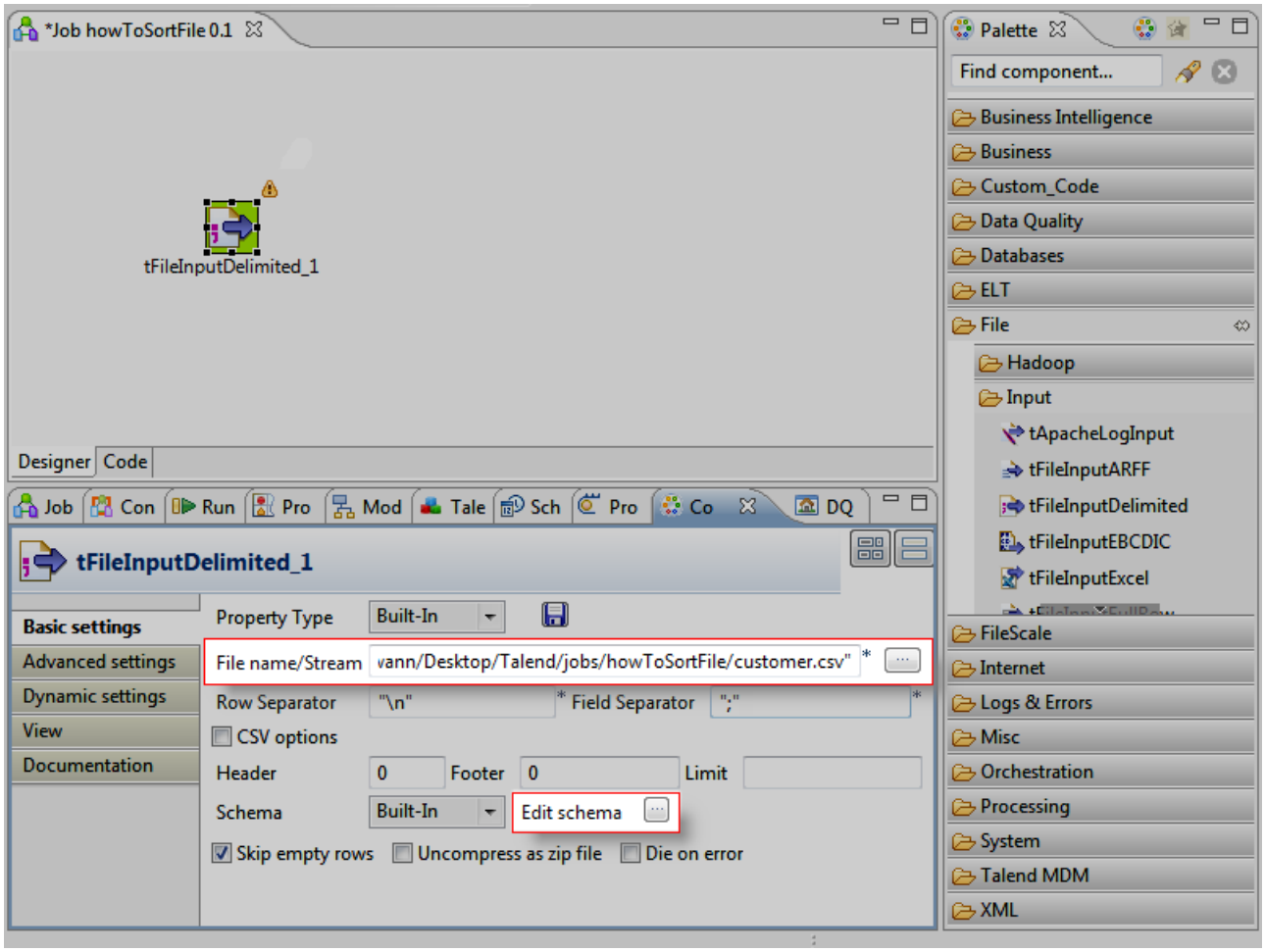
Dans le Job Designer :

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.

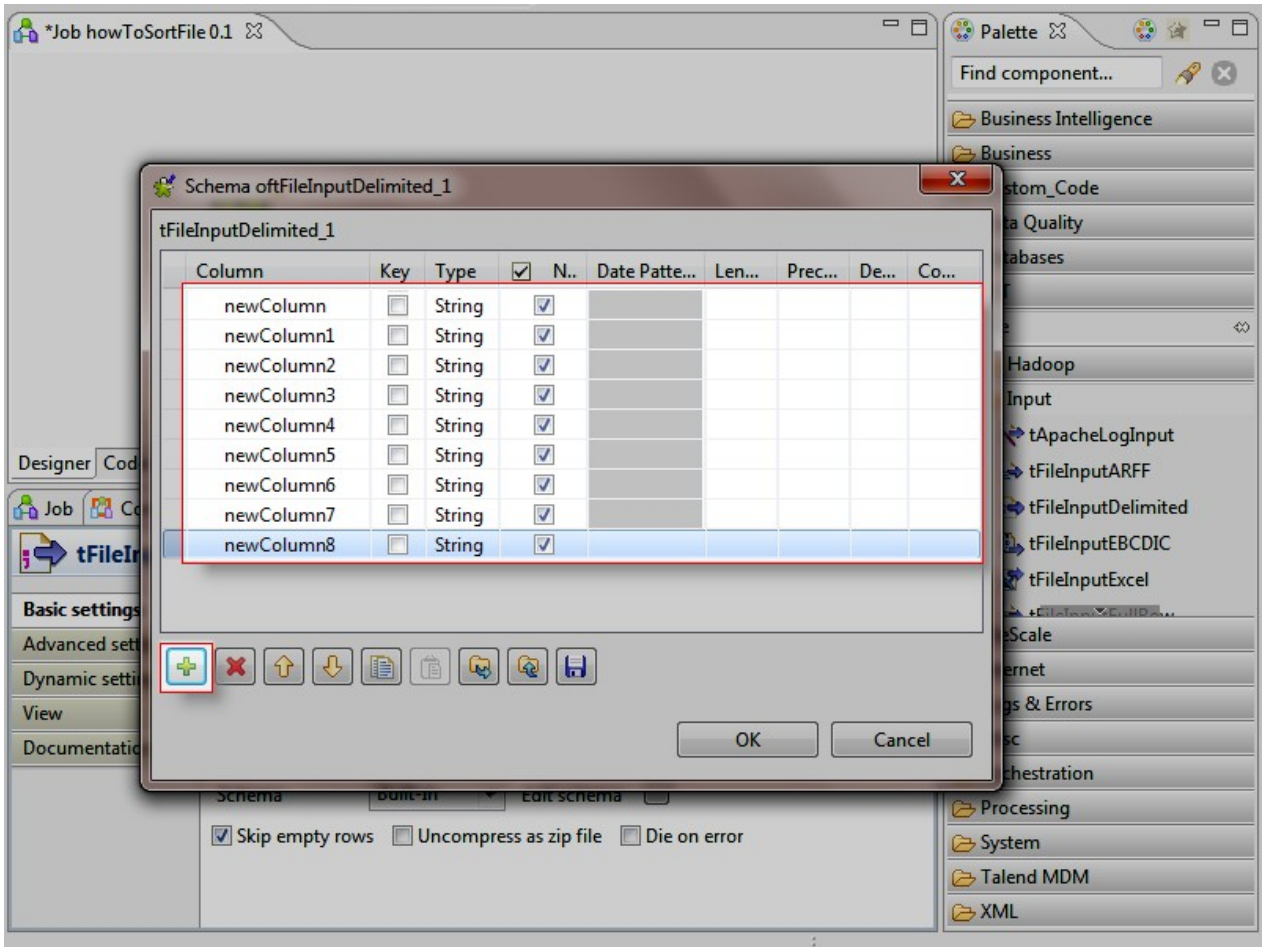
Pour décrire la structure du fichier, cliquez sur le bouton **[...]** situé à coté du champ **Edit schema** pour ouvrir l'assistant "Schema of tFileInputDelimited_1".



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.



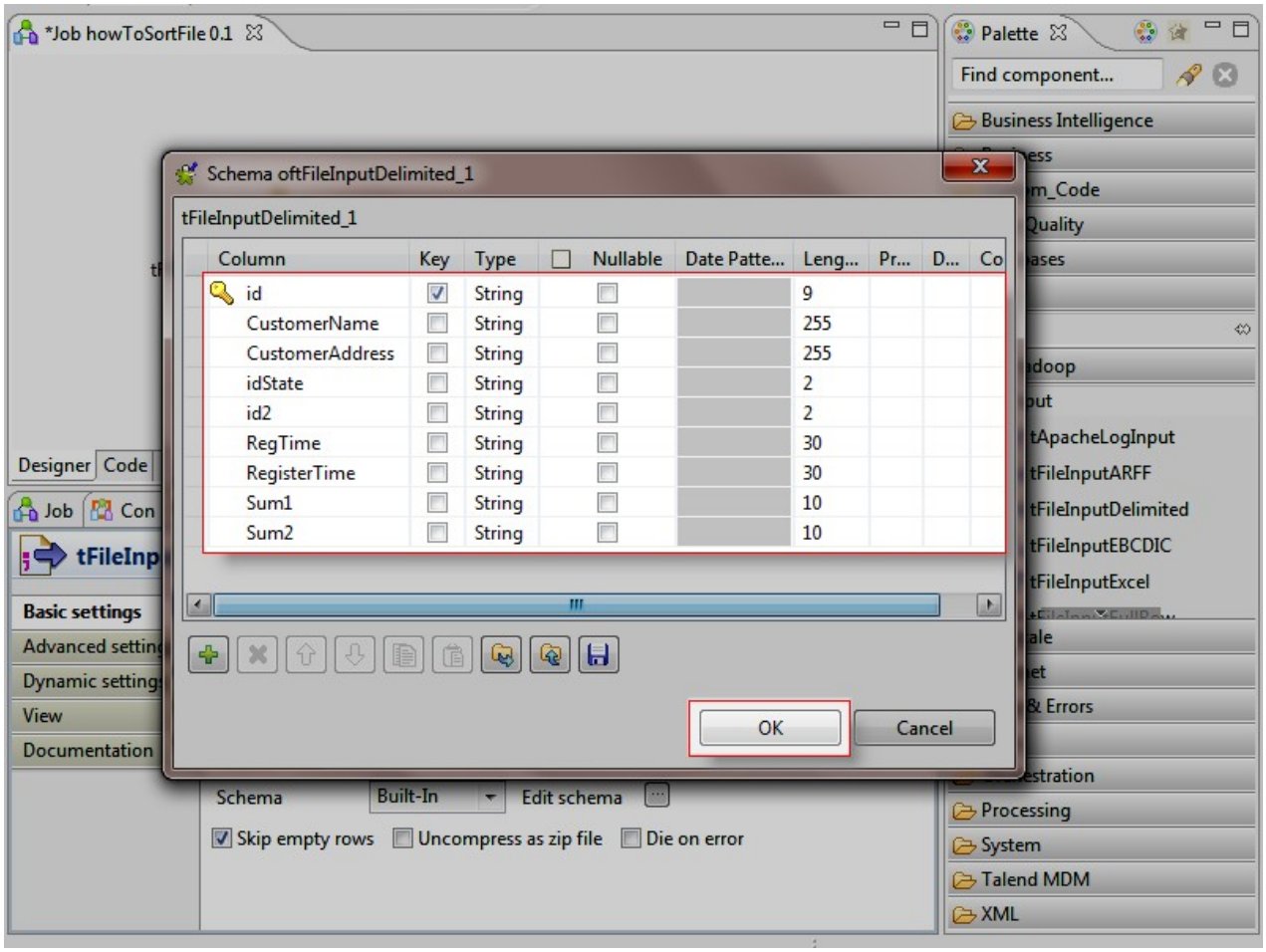
Dans l'assistant Schema of tFileInputDelimited_1 :

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.

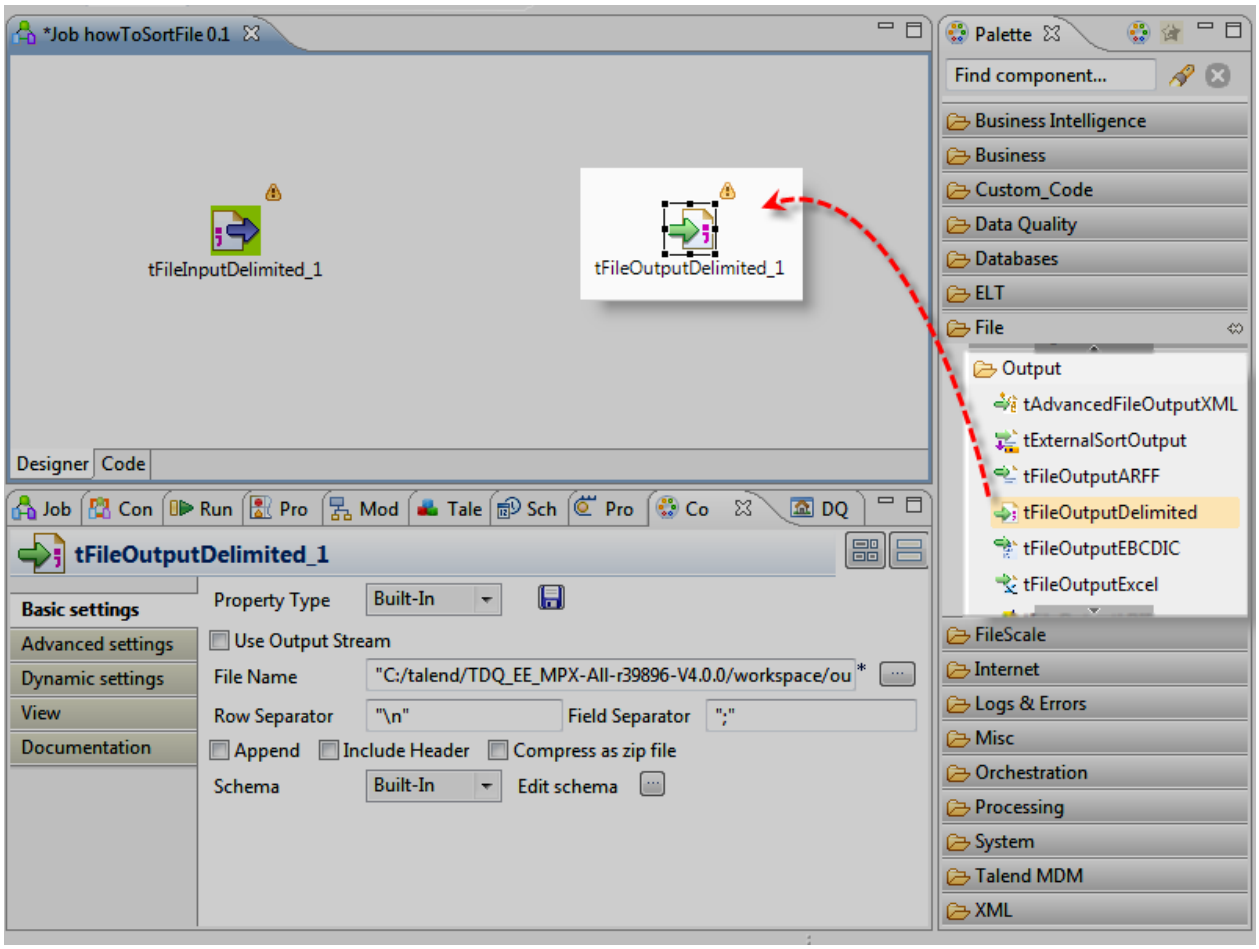


Paramétrer le connecteur d'écriture de fichier

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 :

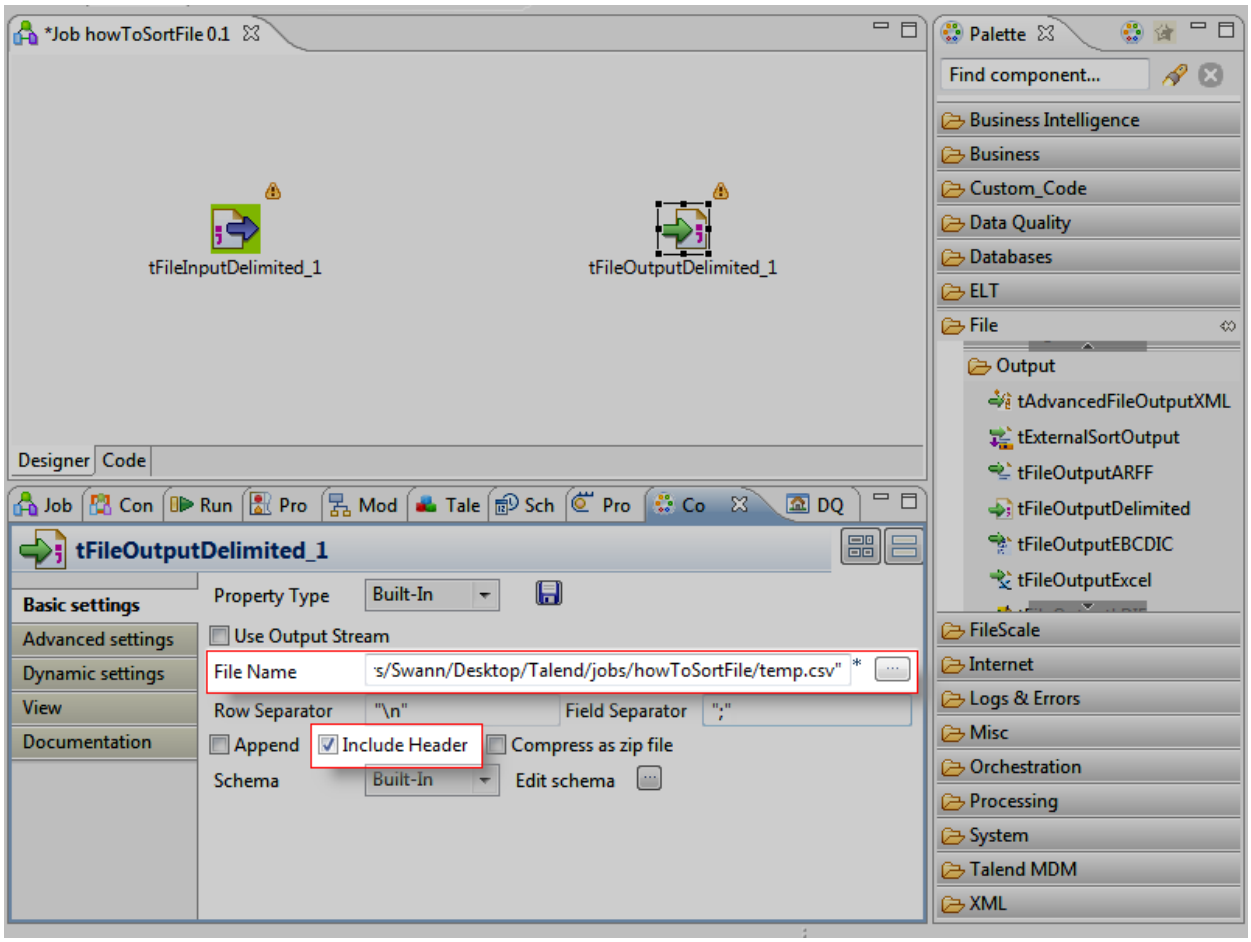
Pour paramétrer les propriétés du **tFileOutputDelimited**, double-cliquez dessus et la vue **Component** correspondante apparaît.

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.

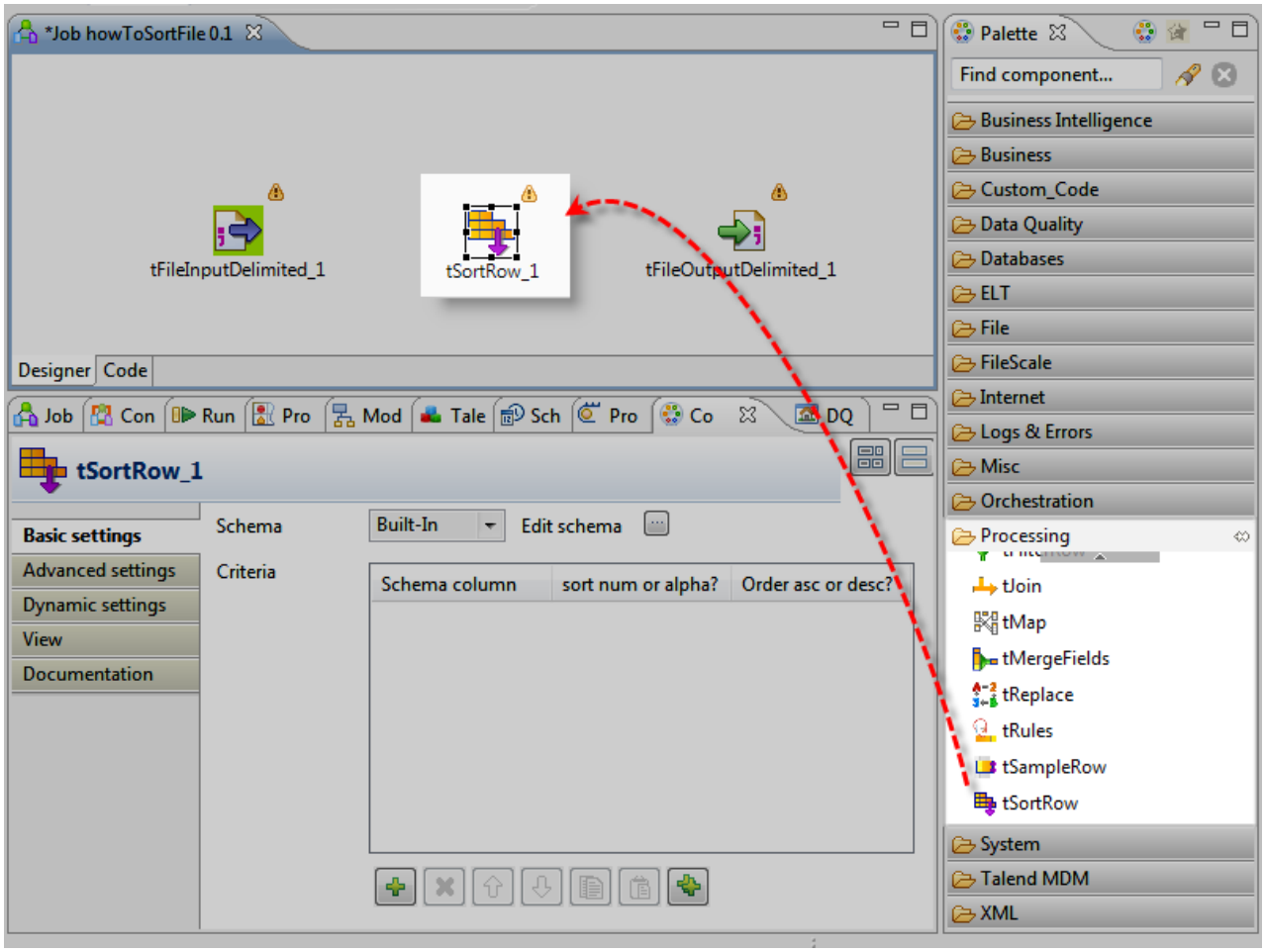


Définir le composant de transformation et relier les composants entre eux

Dans la Palette à droite :

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

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

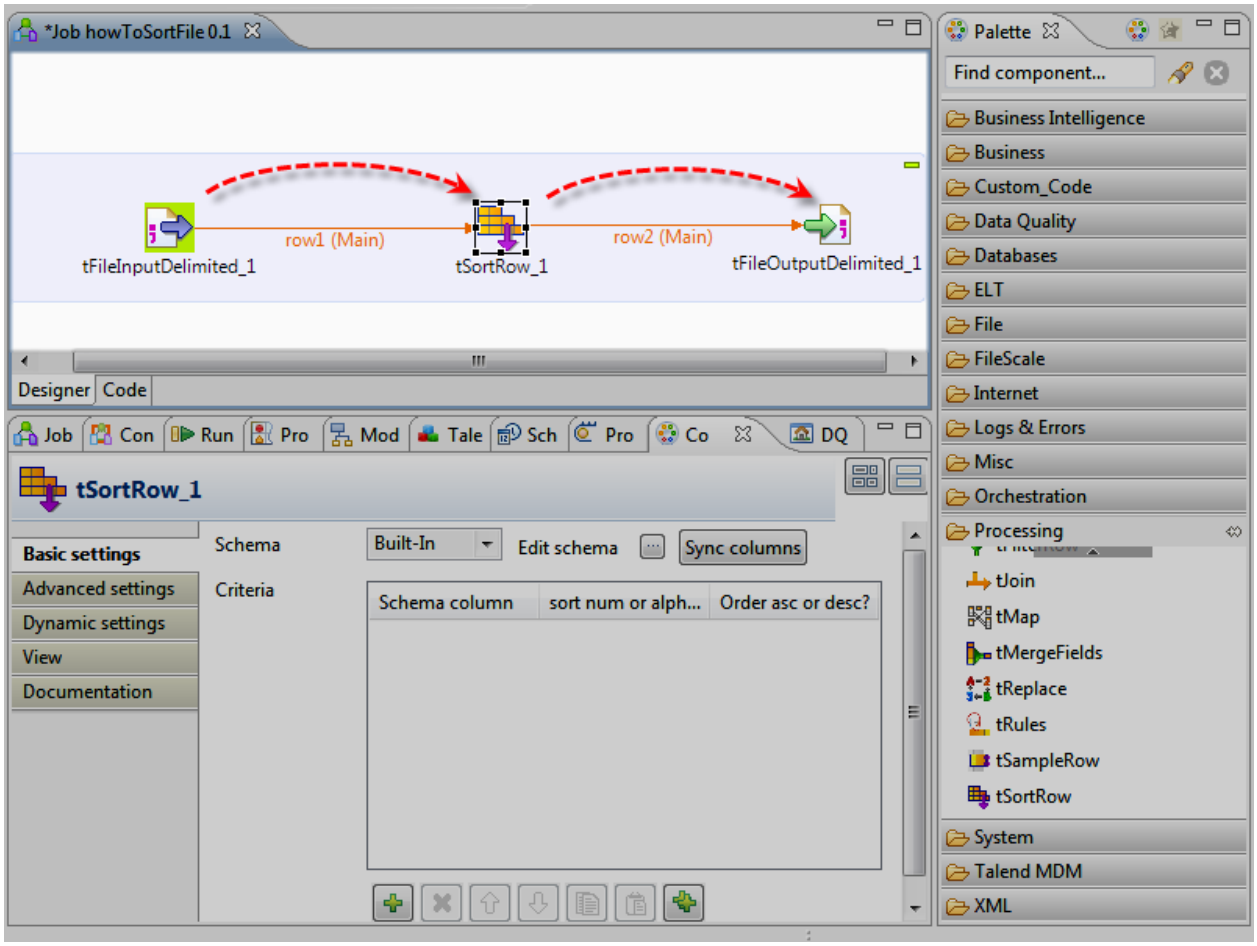


Dans le Job designer :

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.



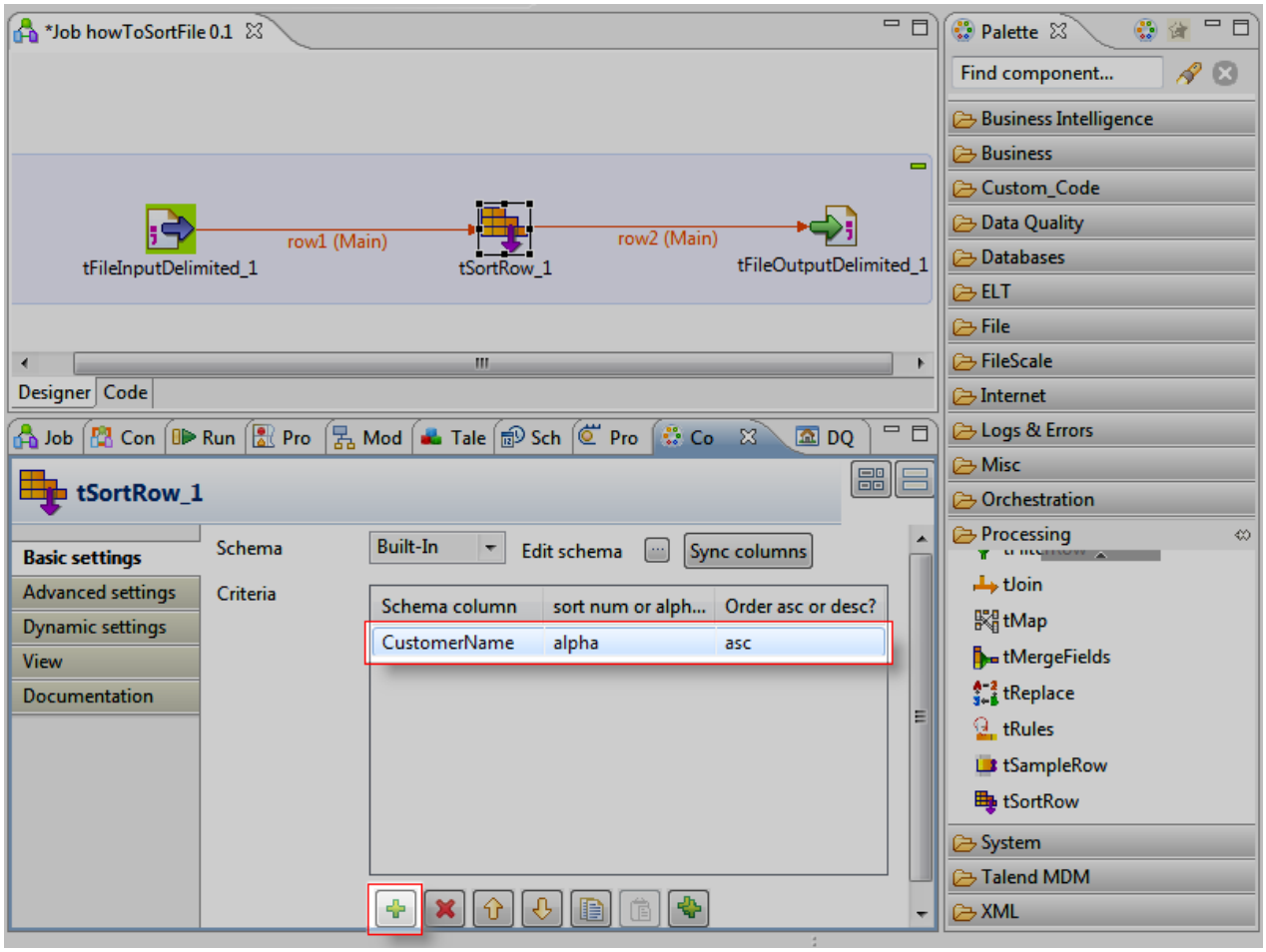
Dans le Job designer :

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.

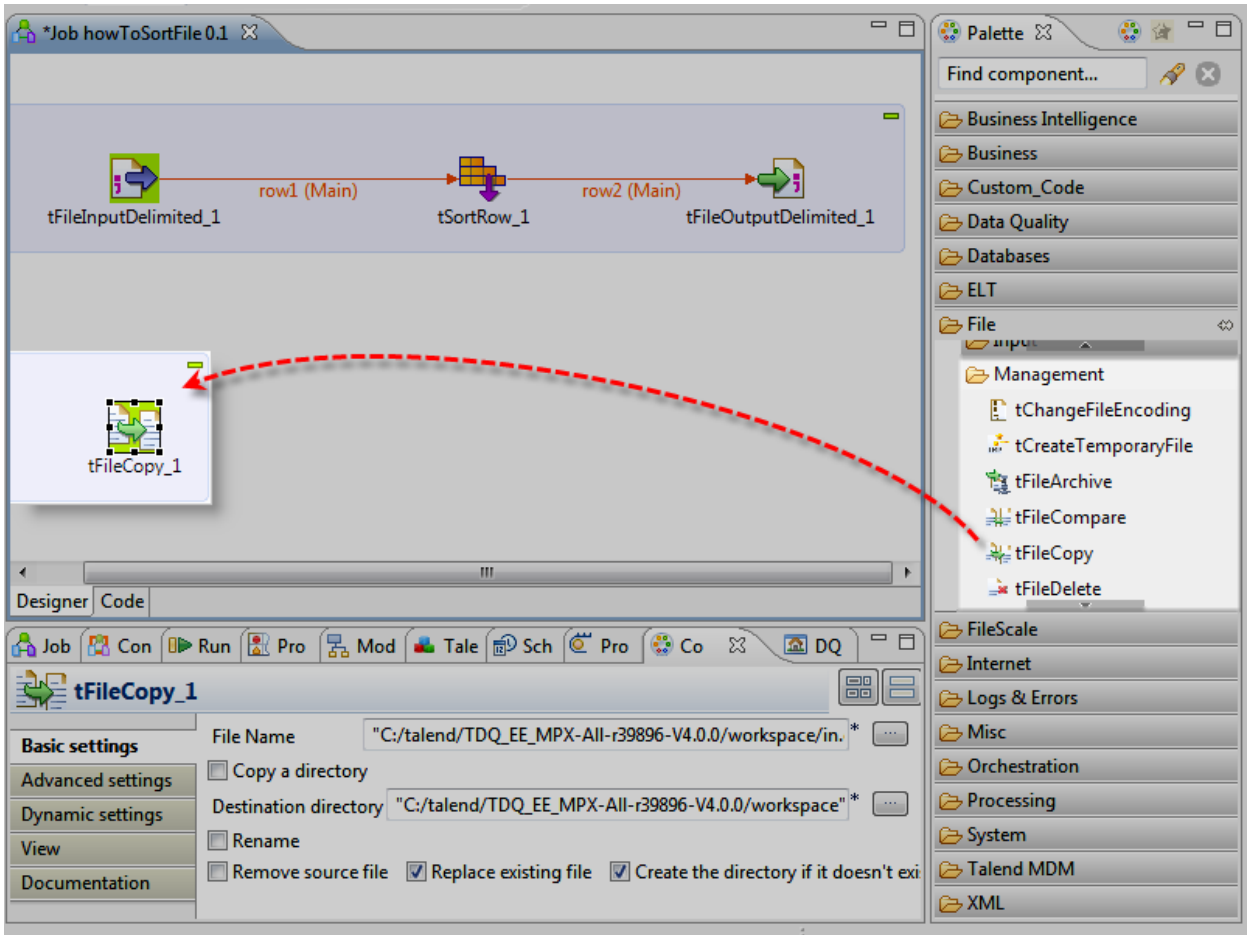
L'objectif étant de trier le fichier original et non d'en créer un nouveau, il nous reste à remplacer le fichier original par ce nouveau fichier.

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 (Gestion)**.

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



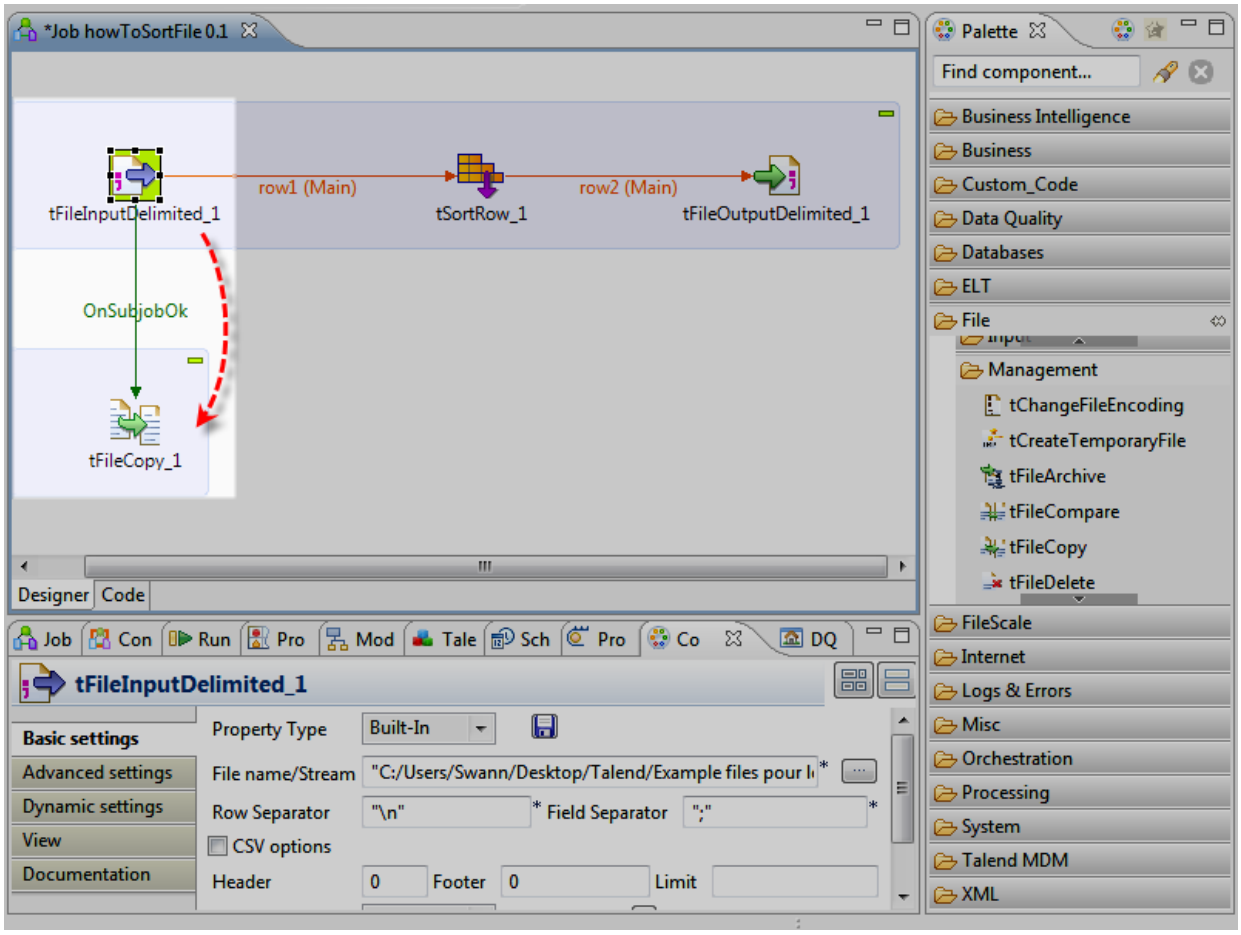
Dans le Job designer :

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.



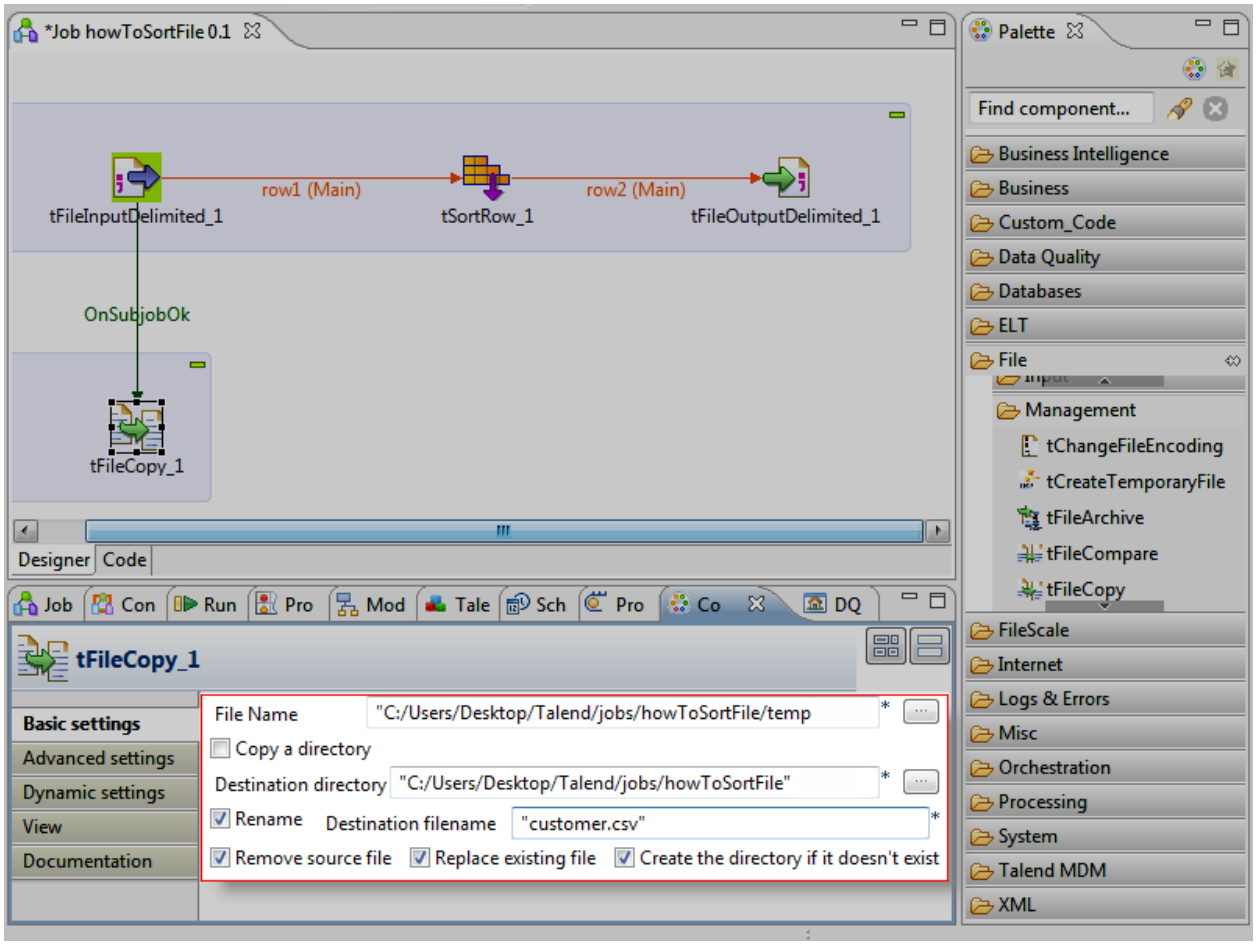
Dans la vue Component :

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**.



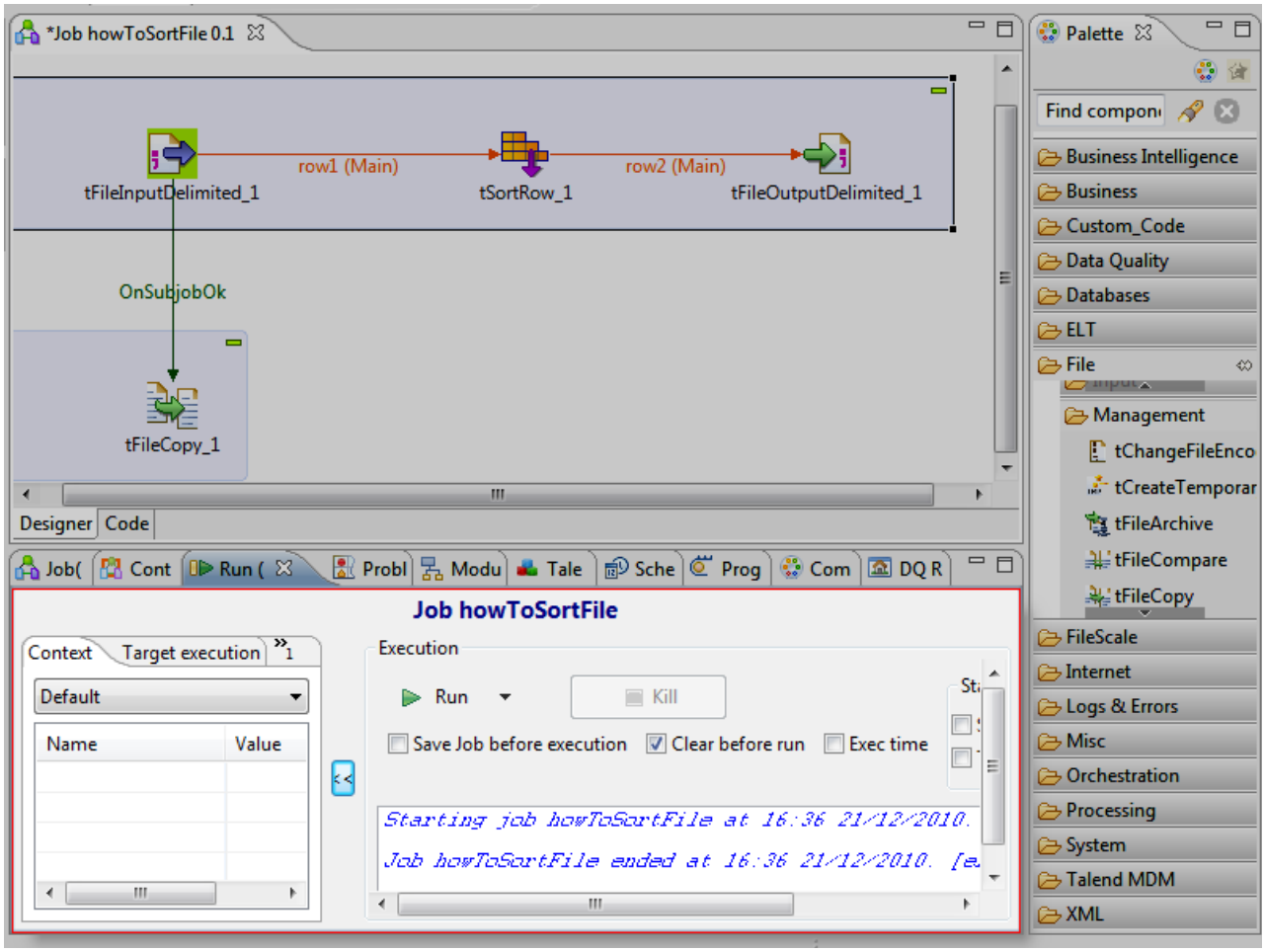
Dans le Job Designer :

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 !

Dans le Job Designer :

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

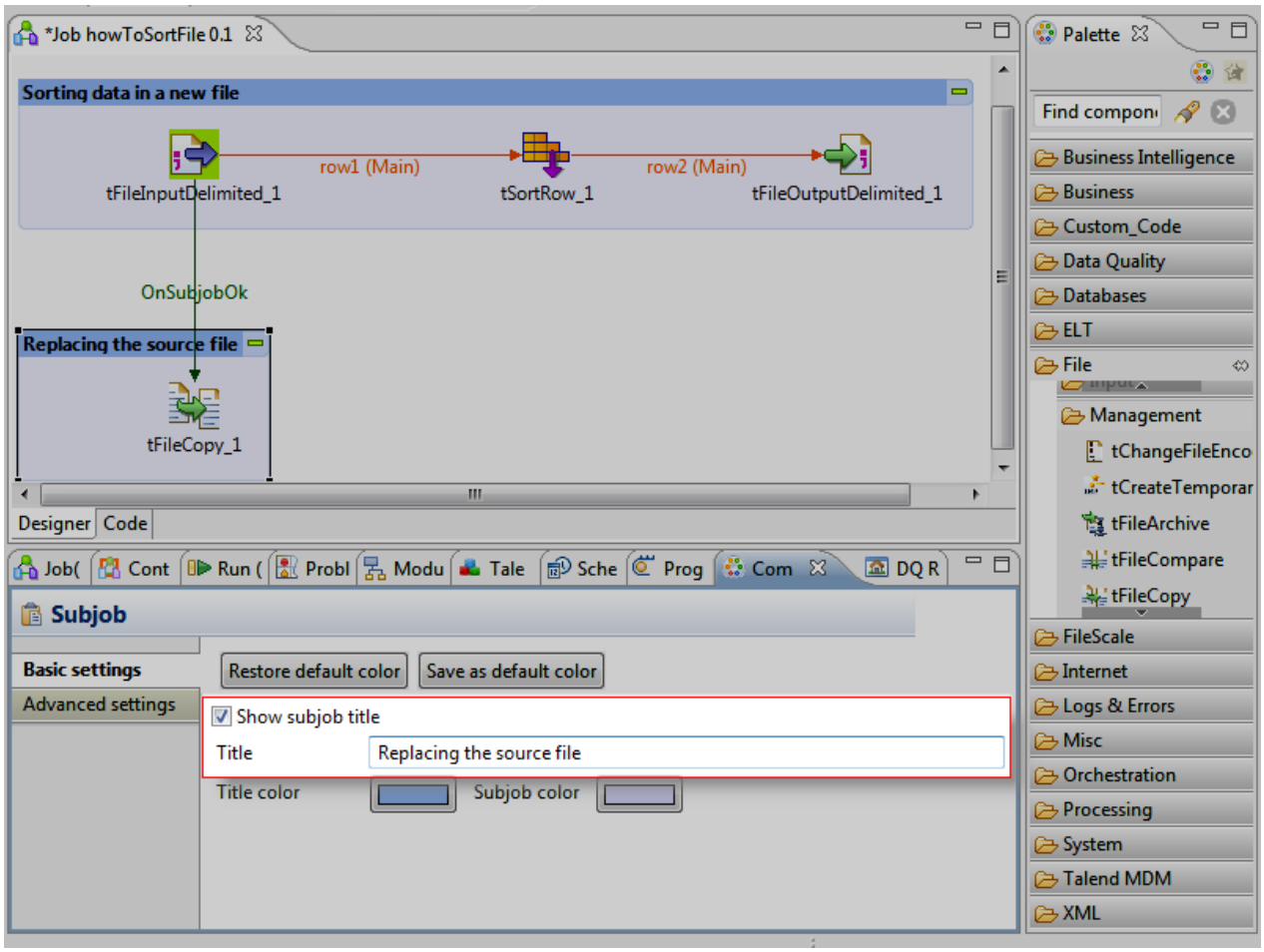
Cliquez sur la zone bleue entourant votre premier sous-job.

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.



Ce scénario est maintenant fini.

Le Job fonctionne et est documenté.