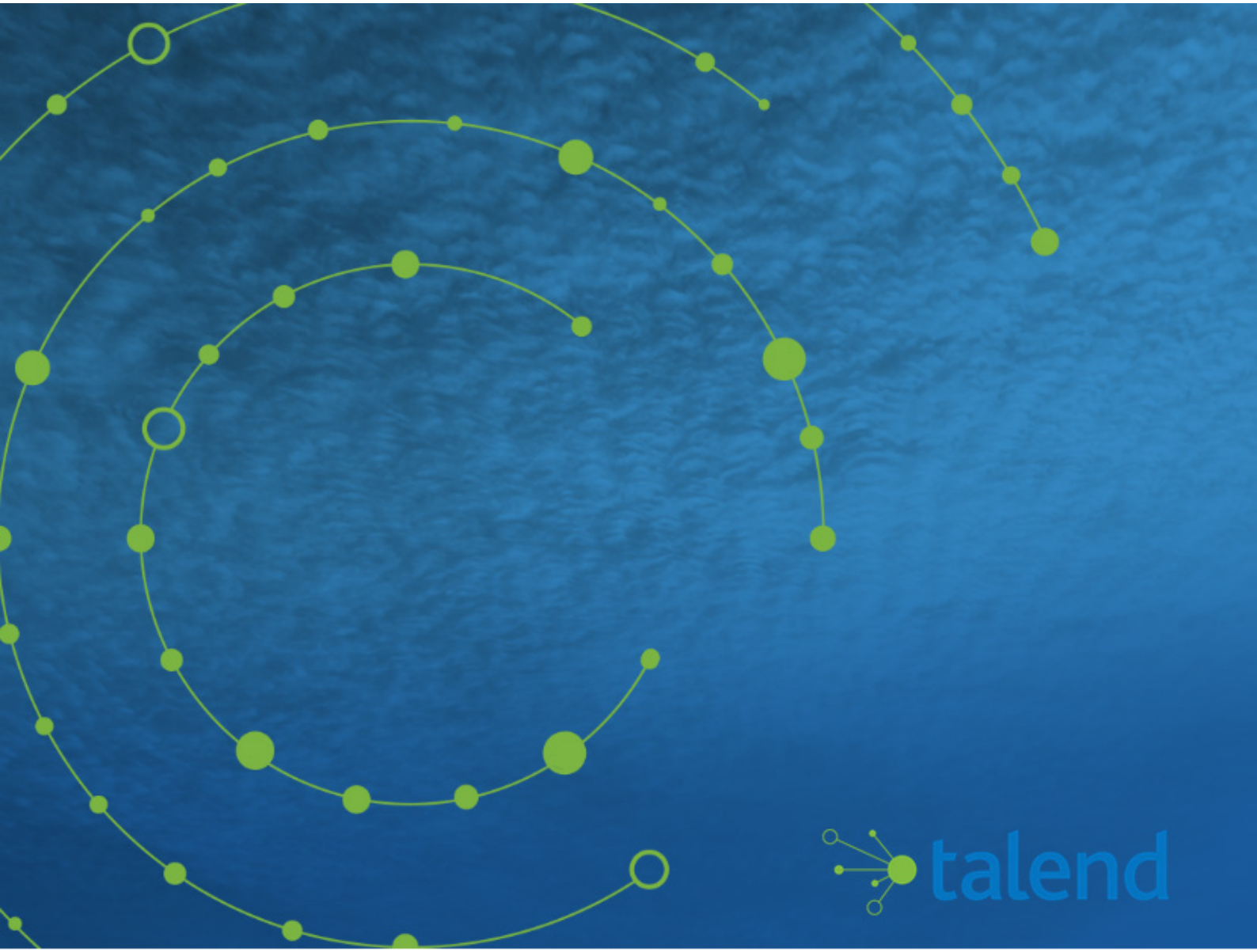


Guide de prise en main de Talend Open Studio for Data Integration

7.0.1

Table des matières

Copyleft

Convient à la version 7.0.1. Annule et remplace toute version antérieure de ce guide.

Date de publication : 13 avril 2018

Introduction à Talend Open Studio for Data Integration

Talend fournit des outils de développement et de gestion unifiés pour intégrer et traiter toutes vos données dans un environnement graphique simple à utiliser.

La solution d'intégration de données de Talend permet aux entreprises de gérer des systèmes de plus en plus complexes en proposant des outils ETL pour répondre aux besoins d'analyse et d'intégration opérationnelle et en offrant des fonctionnalités d'industrialisation.

Prérequis à l'utilisation de Talend Open Studio for Data Integration

Ce chapitre vous fournit des informations simples concernant le logiciel et le matériel requis et recommandés pour prendre en main votre Talend Open Studio for Data Integration.

Il vous guide également à travers les étapes d'installation et de configuration des outils tiers requis et recommandés :

Recommandations relatives à la mémoire

Pour optimiser l'utilisation des produits Talend, référez-vous aux recommandations de mémoire et espace disque ci-dessous :

Utilisation de la mémoire	3 Go minimum, 4 Go recommandés
Utilisation du disque	3 Go

Recommandations logicielles

Pour optimiser l'utilisation des produits Talend, référez-vous aux recommandations système et logicielles ci-dessous :

Logiciels requis

- Systèmes d'exploitation pour le Studio Talend :

Type de support	**Système d'exploitation (64 bits seulement)**
Recommandé	Ubuntu 16.04 LTS
Recommandé	Microsoft Windows 10
Supporté	Apple macOS 10.13/High Sierra
	Apple macOS 10.12/Sierra
	Apple macOS 10.13/High Sierra
	Apple OS X 10.11/El Capitan

- Java 8 JRE Oracle. Consultez Installation de Java à la page 6.
- Une base de données MySQL installée et configurée, avec une base de données `gettingstarted`.

Logiciel facultatif

- 7-Zip. Consultez Installation de 7-Zip (Windows) à la page 7.

Installation de Java

Pour utiliser votre produit Talend, vous avez besoin d'une JRE Oracle (Oracle Java Runtime Environment) installée sur votre ordinateur.

Procédure

1. Dans la page Java SE Downloads (en anglais), sous **Java Platform, Standard Edition**, cliquez sur **JRE Download**.
2. Dans la page **Java SE Runtime Environment 8 Downloads**, sélectionnez le bouton radio **Accept License Agreement**.
3. Sélectionnez le téléchargement correspondant à votre système d'exploitation.
4. Suivez les étapes d'installation de Java proposées par l'assistant Oracle.

Résultats

Lorsque Java est installé sur votre ordinateur, vous devez configurer la variable d'environnement `JAVA_HOME`. Pour plus d'informations, consultez :

- Configuration des variables d'environnement Java sous Windows à la page 6.
- Configuration des variables d'environnement Java sous Linux à la page 7.

Configuration des variables d'environnement Java sous Windows

Avant d'installer votre produit Talend, vous devez configurer les variables d'environnement JAVA_HOME et Path :

Procédure

1. Dans le menu **Démarrer** de votre ordinateur, cliquez-droit sur **Ordinateur** et sélectionnez **Properties**.
2. Dans la fenêtre **Control Panel Home**, cliquez sur **Advanced system settings**.
3. Dans la fenêtre **System Properties**, cliquez sur **Environment Variables...**.
4. Sous **System Variables**, cliquez sur **New...** pour créer une variable. Nommez la variable `JAVA_HOME`, saisissez le chemin d'accès à votre JRE 8 Java, puis cliquez sur **OK**.

   Exemple de chemin vers la JRE par défaut : `C:\Program Files\Java\jre1.8.0_77`.
5. Sous **System Variables**, sélectionnez la variable **Path** et cliquez sur **Edit...** pour ajouter la variable `JAVA_HOME` précédemment définie à la fin de la variable d'environnement `Path`, en les séparant par un point-virgule.

   Exemple : `<PathVariable>;%JAVA_HOME%\bin`.

Configuration des variables d'environnement Java sous Linux

Avant d'installer votre produit Talend, vous devez configurer les variables d'environnement JAVA_HOME et Path.

Procédure

1. Trouvez le répertoire d'installation de la JRE.

   Exemple : `/usr/lib/jvm/jre1.8.0_65`
2. Spécifiez-le dans la variable d'environnement `JAVA_HOME`.

   Exemple :

   ```
   export JAVA_HOME=/usr/lib/jvm/jre1.8.0_65
   export PATH=$JAVA_HOME/bin:$PATH
   ```
3. Ajoutez ces lignes à la fin des profils utilisateurs dans le fichier `~/.profile` ou, en tant que super-utilisateur, à la fin des profils globaux dans le fichier `/etc/profile`.
4. Connectez-vous à nouveau.

Installation de 7-Zip (Windows)

Talend recommande d'installer 7-Zip et de l'utiliser pour extraire les fichiers d'installation : http://www.spiroo.be/7zip/.

Procédure

1. Téléchargez l'installeur de 7-Zip correspondant à votre système d'exploitation.
2. Naviguez dans vos dossiers locaux, trouvez le fichier .exe de 7-Zip et double-cliquez dessus pour l'installer.

Résultats

Le téléchargement démarre automatiquement.

Téléchargement et installation de Talend Open Studio for Data Integration

Talend Open Studio for Data Integration est simple à installer. Après l'avoir téléchargé depuis le site Web de Talend, un simple dézippage permet de l'installer sur votre ordinateur.

Ce chapitre vous fournit les informations de base relatives au téléchargement et à l'installation.

Téléchargement de Talend Open Studio for Data Integration

Talend Open Studio for Data Integration est un produit open source libre que vous pouvez télécharger directement depuis le site Web de Talend.

Procédure

1. Allez à la page de téléchargement de Talend Open Studio for Data Integration.
2. Cliquez sur **TÉLÉCHARGER L'OUTIL LIBRE**.

Résultats

Le téléchargement démarre automatiquement.

Installation de Talend Open Studio for Data Integration

L'installation s'effectue en dézippant le fichier .zip précédemment téléchargé.

Vous pouvez faire ceci en utilisant :

- 7-Zip (recommandé sous Windows) : Extraire via 7-Zip (recommandé pour Windows) à la page 8.
- le dézippeur par défaut de Windows : Extraire via l'outil de dézippage Windows par défaut à la page 9.
- le dézippeur par défaut de Linux (pour un système d'exploitation basé Linux) : Extraire via l'outil de dézippage Windows par défaut à la page 9.

Extraire via 7-Zip (recommandé pour Windows)

Sous Windows, Talend vous recommande d'installer 7-Zip et de l'utiliser pour extraire des fichiers. Pour plus d'informations, consultez Installation de 7-Zip (Windows) à la page 7.

Pour installer le Studio, suivez les étapes suivantes :

Procédure

1. Naviguez dans vos dossiers locaux, trouvez le fichier .zip et déplacez-le à un autre emplacement, avec un chemin d'accès aussi court que possible et sans caractère d'espace.

   Exemple : `C:/Talend/`
2. Dézippez-le en cliquant-droit sur le fichier compressé et sélectionnez **7-Zip** > **Extract Here**.

Extraire via l'outil de dézippage Windows par défaut

Si vous ne souhaitez pas utiliser 7-Zip, vous pouvez utiliser l'outil de dézippage par défaut de Windows :

Procédure

1. Dézippez-le en cliquant-droit sur le fichier compressé et sélectionnez **Extract All**.
2. Cliquez sur **Browse** et naviguez jusqu'au disque C:.
3. Sélectionnez **Make new folder** et nommez le dossier `Talend`. Cliquez sur **OK**.
4. Cliquez sur **Extract** pour commencer l'installation.

Extraire via l'outil de dézippage Linux

Pour installer le Studio, suivez les étapes ci-dessous :

Procédure

1. Naviguez dans vos dossiers locaux, trouvez le fichier .zip et déplacez-le à un autre emplacement, avec un chemin d'accès aussi court que possible, sans caractère d'espace.

   Exemple : `home/user/talend/`
2. Dézippez-le en cliquant-droit sur le fichier compressé et sélectionnez **Extract Here**.

Configuration de votre produit Talend

Ce chapitre prend l'exemple d'une entreprise fournissant des services de locations de films et de streaming de vidéos. Il vous explique comment une telle entreprise peut tirer parti de Talend Open Studio for Data Integration.

Démarrage du Studio pour la première fois

Le répertoire d'installation du Studio contient des fichiers binaires pour différentes plateformes, notamment Mac OS X et Linux/Unix.

Pour ouvrir le Studio Talend pour la première fois, procédez comme suit :

Procédure

1. Double-cliquez sur le fichier exécutable correspondant à votre système d'exploitation, par exemple :
   - `TOS_*-win-x86_64.exe`, sous Windows.
   - `TOS_*-linux-gtk-x86_64`, sous Linux.
   - `TOS_*-macosx-cocoa.app`, sous Mac.
2. Dans la fenêtre **User License Agreement** qui s'ouvre, lisez et acceptez les termes de la licence pour procéder aux étapes suivantes.

Connexion au Studio

Pour vous connecter au Studio Talend pour la première fois, procédez comme suit :

Procédure

1. Dans la fenêtre de login du Studio Talend, sélectionnez l'option **Create a new project**, spécifiez le nom du projet : `getting_started` et cliquez sur **Finish** pour créer un nouveau projet local.
2. Selon le produit que vous utilisez, vous voyez s'ouvrir :
   - la visite guidée (Quick Tour). Jouez-la pour obtenir plus d'informations relatives à l'interface du Studio, puis cliquez sur **Stop** pour la terminer.
   - la page de bienvenue (**Welcome**). Suivez les liens pour obtenir plus d'informations concernant le Studio et cliquez sur **Start Now!** pour fermer la page et continuer l'ouverture du Studio.

   Conseil :

   Une fois votre Studio démarré, vous pouvez également cliquer sur le lien **Videos**, en haut de la fenêtre principale du Studio pour visionner quelques vidéos courtes pouvant vous aider à prendre en main votre Studio Talend. Pour certains systèmes d'exploitation, vous devrez peut-être installer un décodeur/lecteur MP4 pour lire les vidéos.

Résultats

Vous êtes connecté au Studio Talend. Vous devez installer les packages supplémentaires requis pour que le Studio Talend fonctionne correctement.

Installation des packages supplémentaires

Talend vous recommande d'installer des packages supplémentaires, y compris des bibliothèques tierces et les pilotes de bases de données, dès que vous vous connectez à votre Studio Talend, afin de tirer pleinement parti de toutes les fonctionnalités du Studio.

Procédure

1. Lorsque l'assistant **Additional Talend Packages** s'ouvre, installez les packages supplémentaires, en cochant les cases **Required** et **Optional third-party libraries**. Cliquez sur **Finish**.

   Cet assistant s'affiche à chaque fois que vous lancez le studio si des packages supplémentaires sont disponibles à l'installation à moins que vous ne cochiez la case **Do not show this again**. Vous pouvez également afficher cet assistant en sélectionnant **Help** > **Install Additional Packages** dans la barre de menu.

   Pour plus d'informations, consultez la section concernant l'installation de packages supplémentaire s dans le Guide d'installation et de migration de Talend Open Studio for Data Integration.
2. Dans la fenêtre **Download external modules**, cliquez sur le bouton **Accept all** au bas de l'assistant pour accepter toutes les licences des modules externes dans le studio.

   Attendez que toutes les bibliothèques soient installées avant de commencer à utiliser le studio.
3. Si nécessaire, redémarrez votre Studio Talend pour que certains packages supplémentaires soient pris en compte.

Tâches d'intégration de données

Ce chapitre prend l'exemple d'une entreprise fournissant des services de locations de films et de streaming de vidéos. Il vous explique comment une telle entreprise peut tirer parti de Talend Open Studio for Data Integration.

Vous allez utiliser des données relatives à des films et des réalisateurs, tout en apprenant à filtrer les données des films par rapport aux données des réalisateurs et séparer les entrées ayant des informations relatives aux réalisateurs valides de celles ayant de telles informations invalides.

Lire les informations relatives à des films à partir d'un fichier CSV

Les exemples fournis dans ce chapitre présupposent que :

- vous ayez démarré votre Studio Talend et ouvert la perspective **Integration**.
- vous ayez installé toutes les bibliothèques tierces et les pilotes de bases de données requis dans votre Studio Talend.
- vous ayez installé et configuré le logiciel de base de données MySQL et créé une base de données **gettingstarted**.

Au cours de ce scénario, vous allez apprendre à :

- Créer un Job d'intégration de données. Consultez Créer votre premier Job à la page 12 pour plus de détails.
- Ajouter et relier des composants dans un Job d'intégration de données. Consultez Déposer et relier des composants à la page 13 pour plus de détails.
- Créer une métadonnée de fichier dans le **Repository**. Consultez Préparer la métadonnée relative aux films à la page 14 pour plus de détails.
- Configurer et exécuter un Job d'intégration de données. Consultez Configurer et exécuter votre Job à la page 18 pour plus de détails.

Si vous souhaitez reproduire l'exemple décrit dans ce document et utiliser les données d'entrées exactes, vous pouvez télécharger `tos_di_gettingstarted_source_files.zip` depuis l'onglet **Downloads** de la version en ligne de cette page à l'adresse https://help.talend.com, et sauvegarder les fichiers source dans votre répertoire loca `C:\getting_started\input_data\`.

Créer votre premier Job

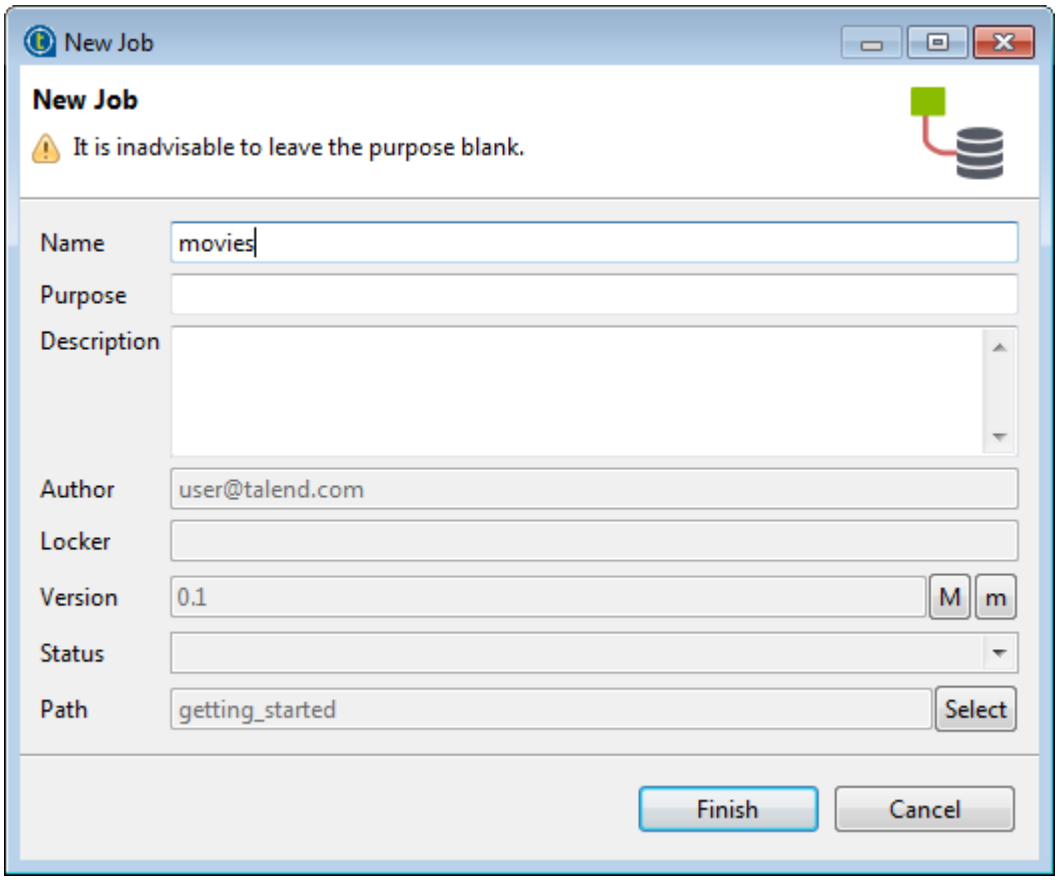
Cette procédure décrit comment créer un dossier de Jobs nommé `getting_started` et un Job nommé `movies`, dans ce dossier.

Procédure

1. Dans la vue **Repository**, cliquez-droit sur le nœud **Job Designs**et sélectionnez **Create folder** dans le menu contextuel.
2. Dans l'assistant **New Folder**, nommez votre dossier de Jobs `getting_started` puis cliquez sur **Finish** pour créer votre dossier.
3. Cliquez-droit sur le dossier **getting_started** et sélectionnez **Create Job** dans le menu contextuel.

4. Dans l'assistant **New Job**, saisissez un nom pour le Job à créer, ainsi que d'autres informations utiles.

   Dans cet exemple, saisissez `movies` dans le champ **Name**.



   Dans cette étape de l'assistant, **Name** est le seul champ obligatoire. Les informations que vous fournissez dans le champ **Description** s'affichent en tant qu'info-bulle lorsque vous passez votre curseur sur le Job dans la vue **Repository** tree view.

5. Cliquez sur **Finish** pour créer votre Job.

   Un Job vide s'ouvre dans le Studio.

Déposer et relier des composants

Cet exemple décrit comment ajouter et relier des composants dans le nouveau Job créé, pour lire un fichier CSV et afficher les données dans la console.

Procédure

1. Déposez un **tFileInputDelimited** et un **tLogRow** de la **Palette** dans l'espace de modélisation graphique.

   Vous pouvez trouver le composant **tFileInputDelimited** dans le groupe **Input** de la famille **File** et le **tLogRow** dans la famille **Logs & Errors**, dans la **Palette**.

2. Cliquez sur le composant **tFileInputDelimited**, une icône représentant un **o** s'affiche, glissez-dépose z l'icône **o** sur le composant **tLogRow**.

   Les deux composants sont reliés via un lien **Row** > **Main**.

Résultats

Vous avez ajouté les composants nécessaires au Job. Dans les étapes suivantes, vous allez préparer la métadonnée requise et configurer le Job.

Préparer la métadonnée relative aux films

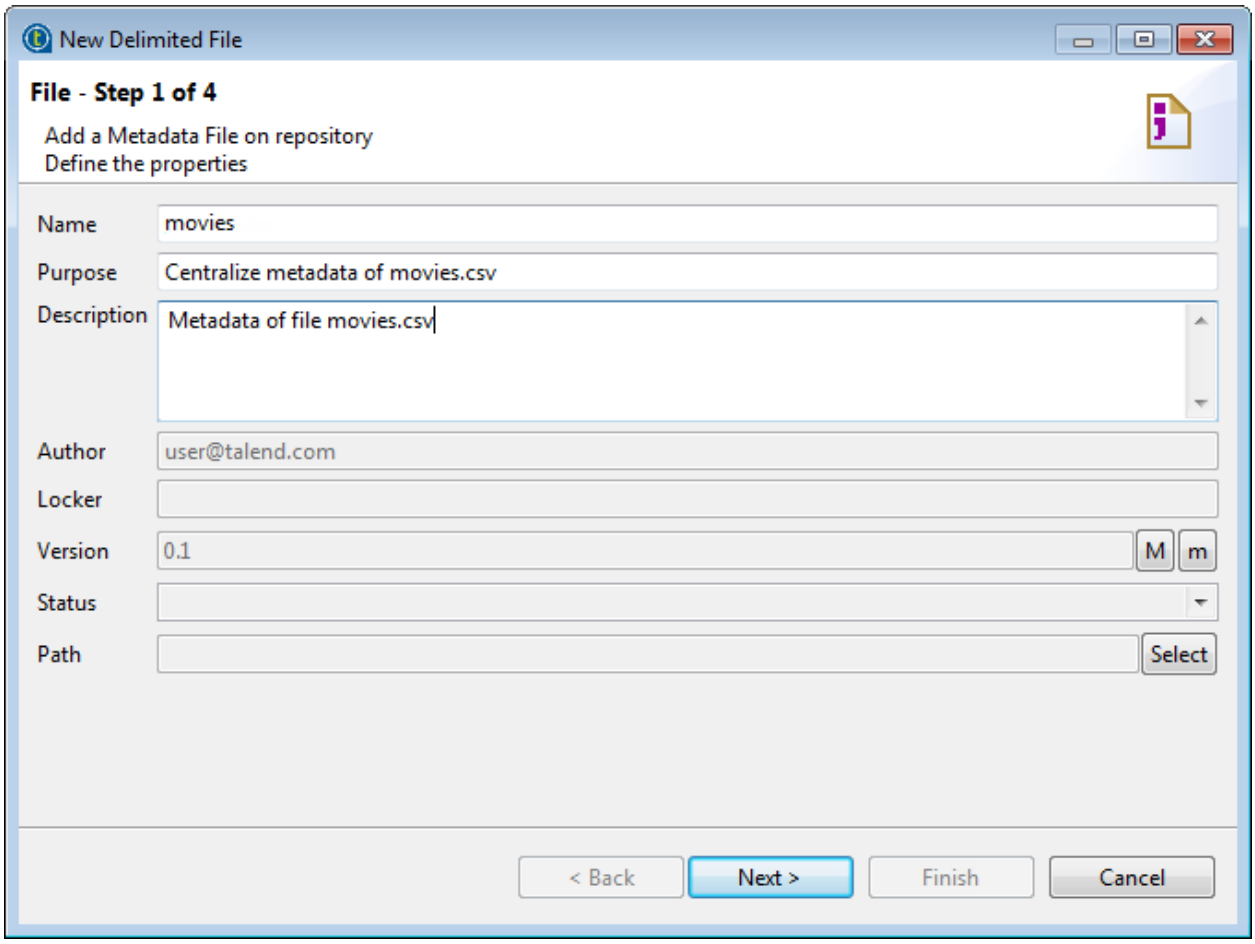
Cette section décrit comment configurer la métadonnée du fichier source `movies.csv` dans le **Repository**. Les métadonnées stockées dans le référentiel peuvent être utilisées dans plusieurs Jobs, vous permettant ainsi de configurer rapidement vos Jobs sans avoir à définir chaque paramètre et schéma manuellement.

Avant de commencer

- Votre fichier source `movies.csv` doit être disponible dans le dossier `C:\getting_started\input_data\`.

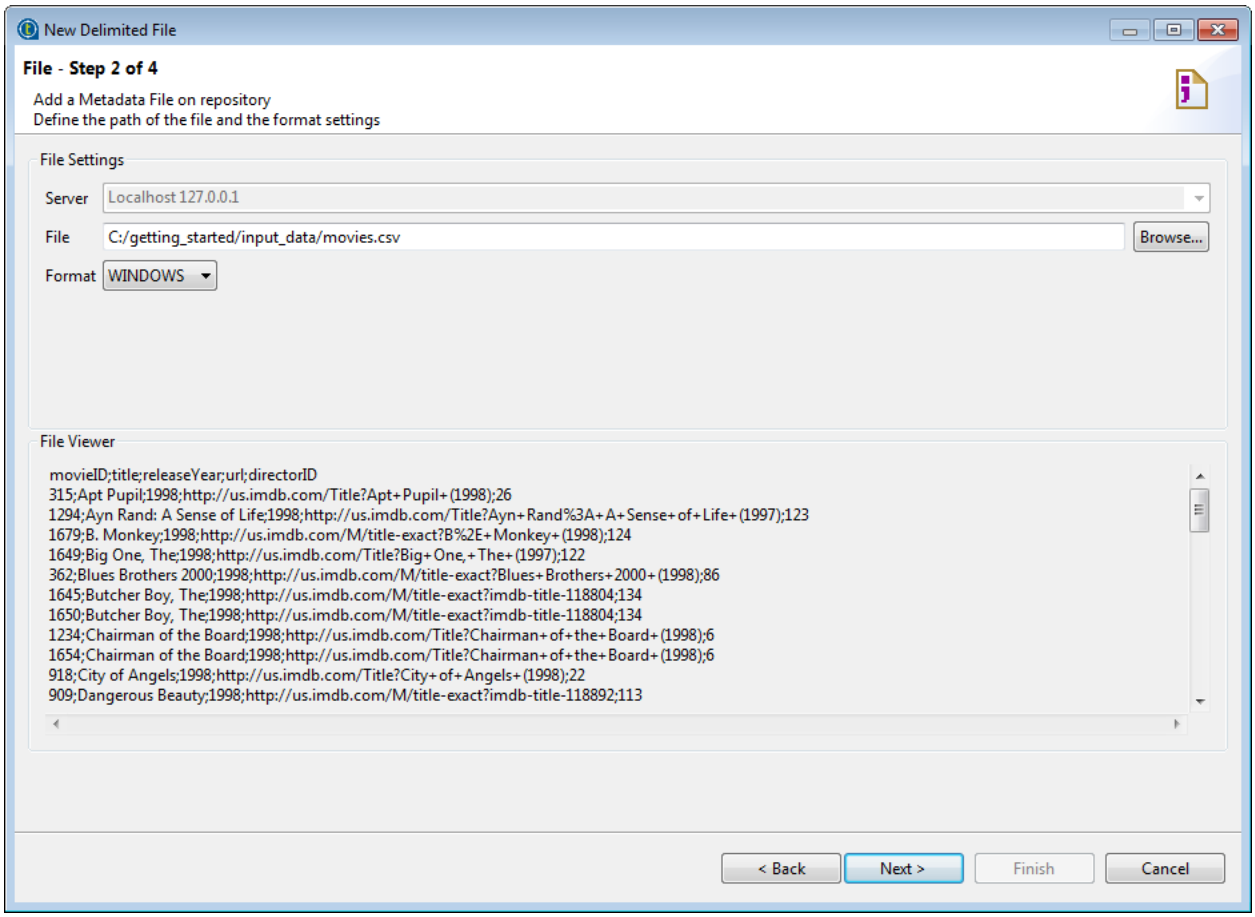
Procédure

1. Dans la vue **Repository**, développez le nœud **Metadata**, cliquez-droit sur **File delimited** et sélectionnez **Create file delimited** dans le menu contextuel pour ouvrir l'assistant **New Delimited File**.
2. Dans l'assistant **New Delimited File**, saisissez un nom pour la métadonnée du fichier, movies, dans cet exemple et d'autres informations utiles permettant de décrire votre métadonnée, puis cliquez **Next** pour passer à l'étape suivante et définir les propriétés générales du fichier.



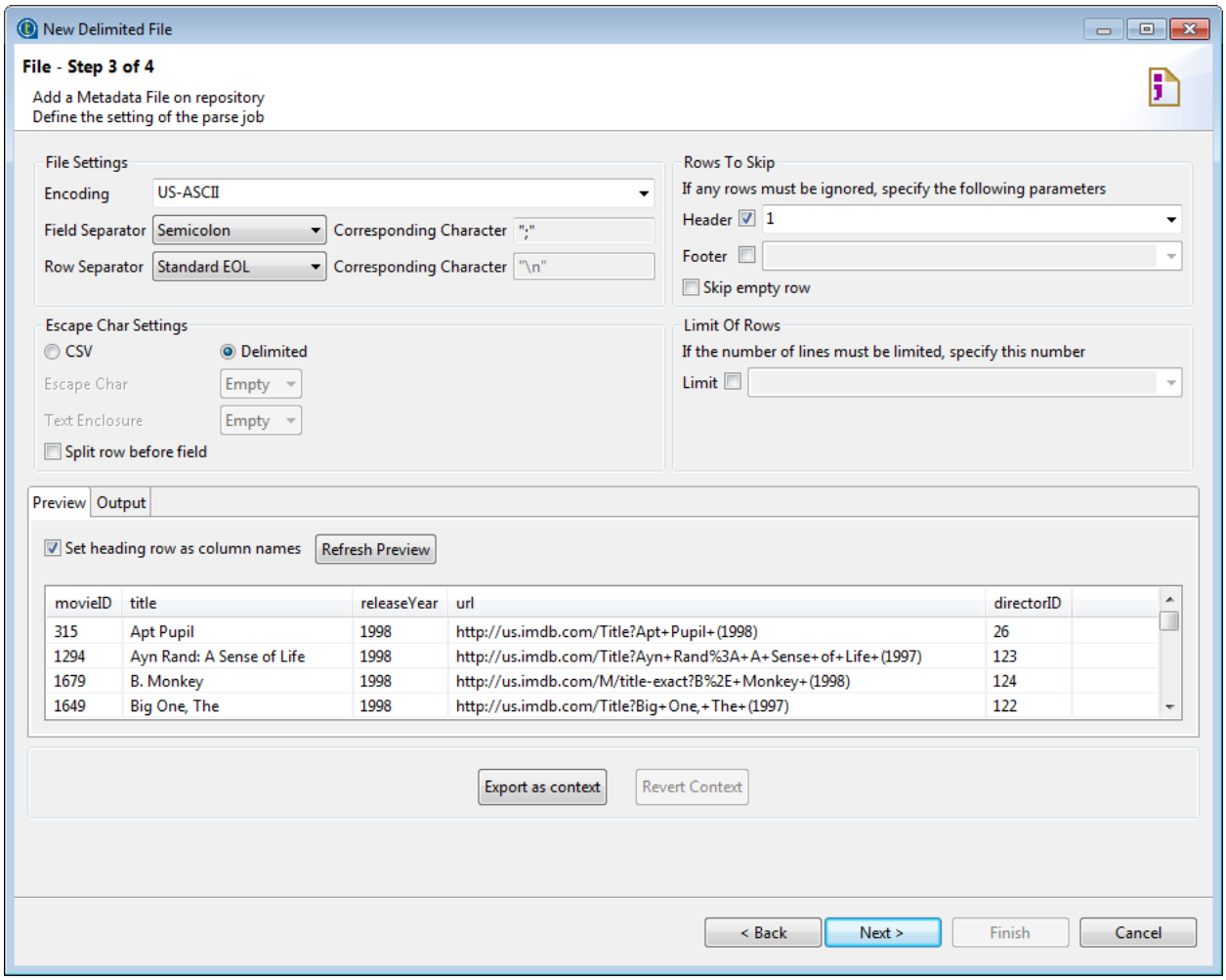
Dans cette étape de l'assistant, **Name** est le seul champ obligatoire. Les informations fournies dans le champ **Description** s'affichent en tant qu'info-bulle lorsque vous placez votre curseur sur la métadonnée.

3. Dans le champ **File**, spécifiez le chemin du fichier source, ou cliquez sur **Browse** pour parcourir votre système jusqu'à ce fichier.



La zone **File Viewer** affiche un aperçu du fichier, vous permettant de vérifier sa cohérence, la présence d'un en-tête et la structure du fichier.

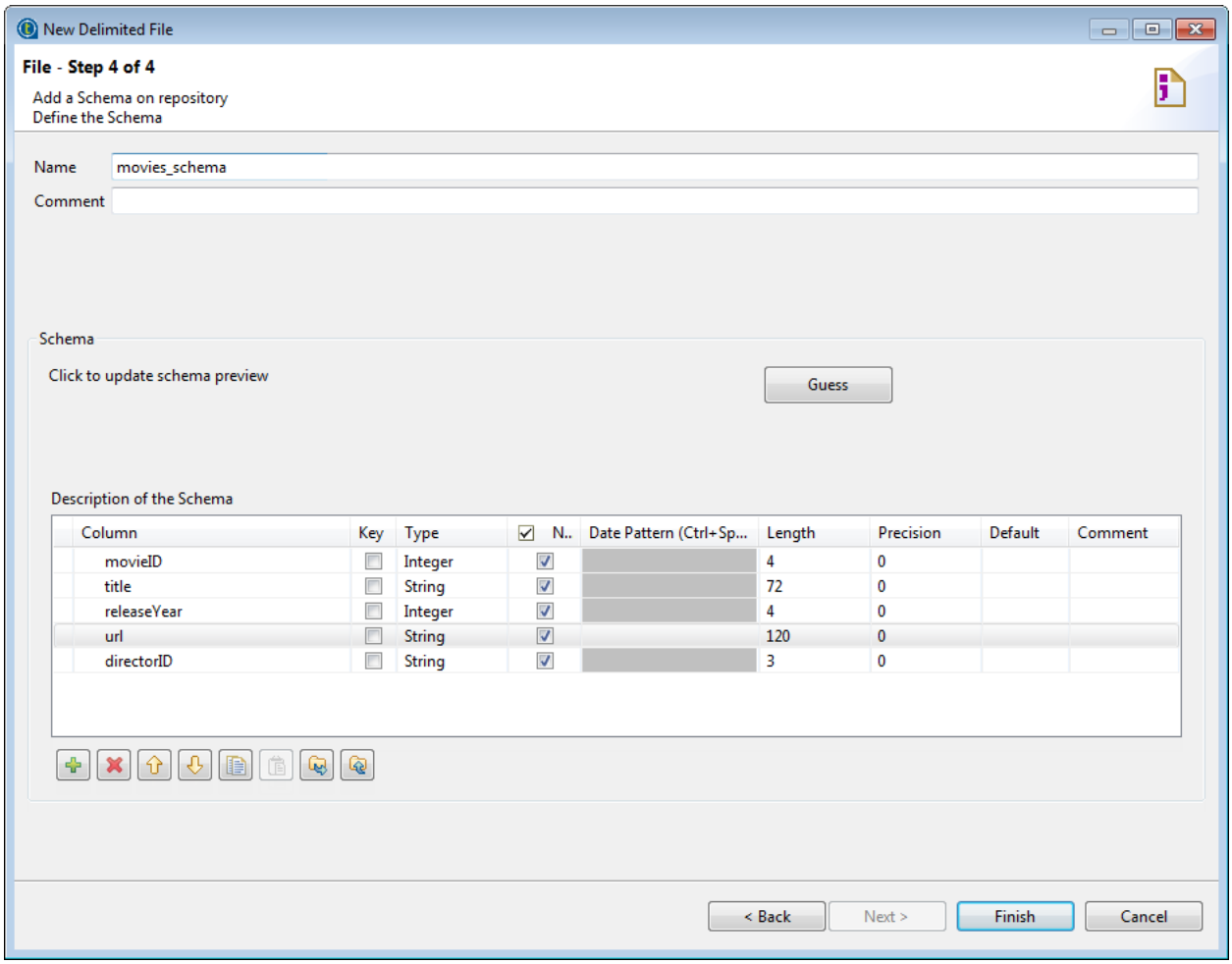
4. Dans la liste **Format**, sélectionnez votre système d'exploitation et cliquez sur **Next** pour parser le fichier.
5. Dans l'onglet **Preview**, cochez la case **Set** heading row as column names pour récupérer les noms de colonnes de la première ligne, puis cliquez sur **Refresh Preview**



La case **Header** de la zone **Rows To Skip** est automatiquement cochée et le nombre de lignes d'en-tête à ignorer est incrémenté de 1.

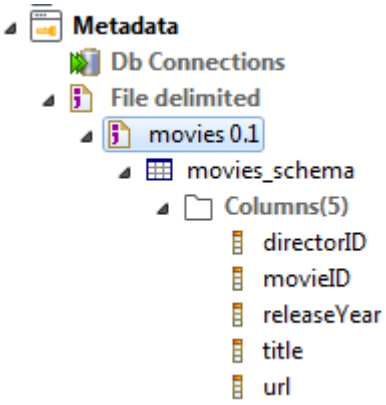
6. Si le fichier contient plusieurs lignes d'en-tête devant être ignorées lors du parsage du fichier, spécifiez le nombre de lignes, dans ce champ, puis cliquez sur **Refresh Preview**.
7. Cliquez sur **Next** pour récupérer le schéma du fichier.
8. Nommez le schéma `movies_schema`, vérifiez-le et modifiez-le selon vos besoins.

   Dans cet exemple, augmentez la valeur dans la colonne Length pour les lignes title et url.



9. Cliquez sur **Finish** pour valider le schéma et fermer l'assistant.

   La métadonnée de fichier créée s'affiche dans la vue **Repository**.



Résultats

La métadonnée du fichier relatif aux films est maintenant prête à être utilisée. Vous allez utiliser cette métadonnée avec votre composant d'entrée lisant le fichier source.

Configurer et exécuter votre Job

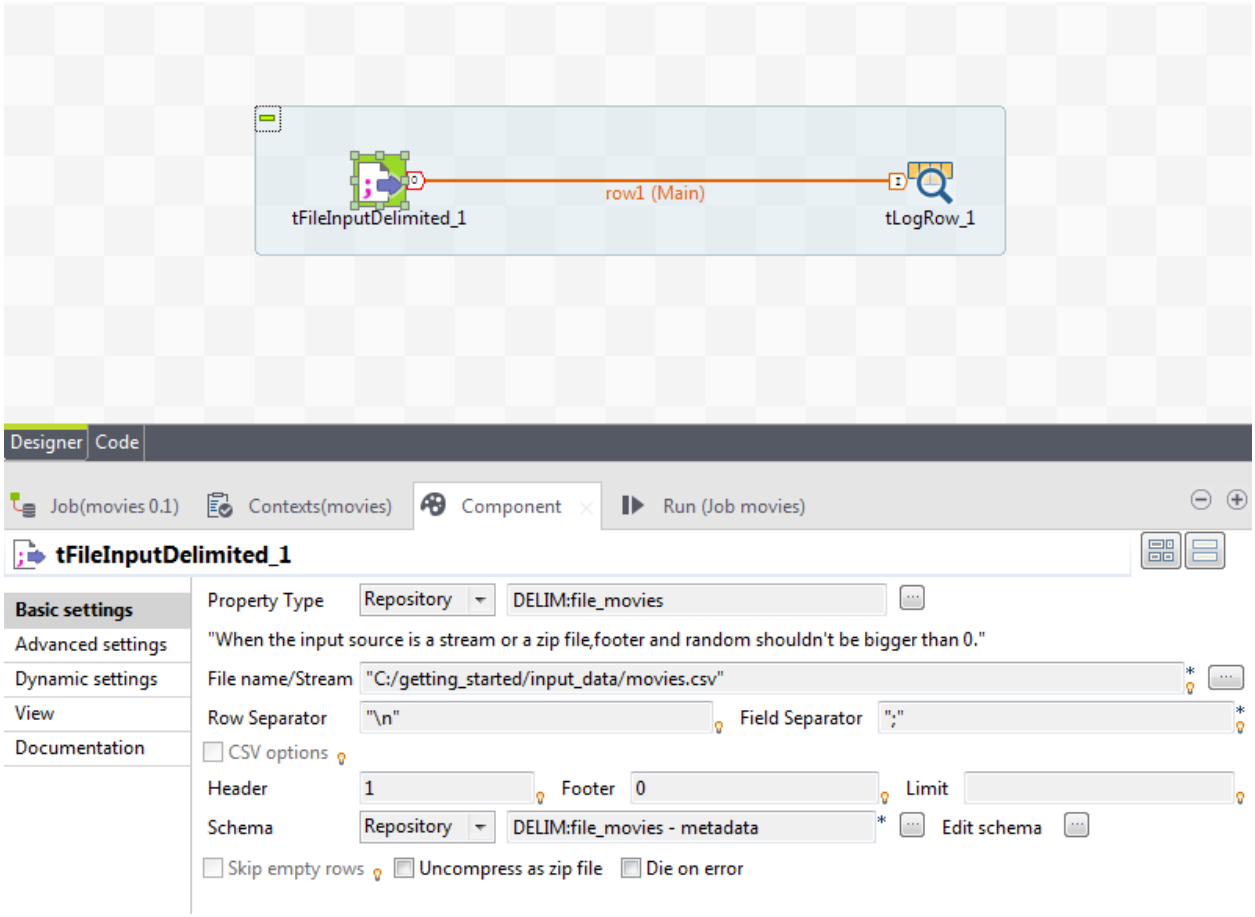
Cet exemple décrit comment configurer les composants à l'aide de la métadonnée créée dans la procédure précédente et comment exécuter votre Job.

Procédure

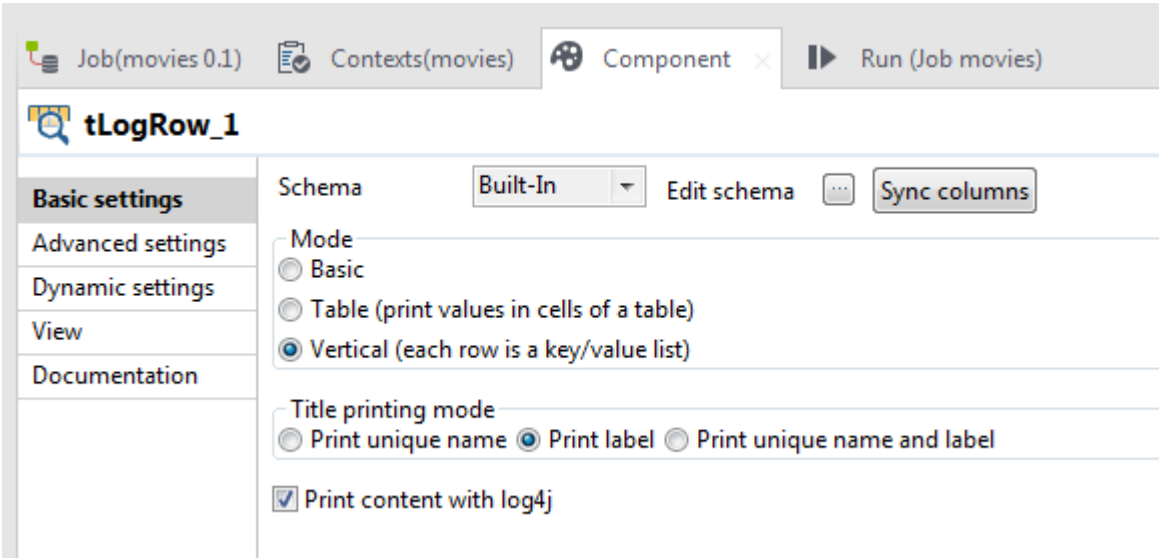
1. Dans la vue **Repository**, double-cliquez sur le Job **movies** pour l'ouvrir dans l'espace de modélisation graphique.

   Vous pouvez ignorer cette étape si le Job est déjà ouvert dans l'espace de modélisation graphique.
2. Dans la vue **Repository**, développez **Metadata** > **File delimited** et glissez-déposez la métadonnée de fichier **movies** ou son schéma **movies_schema** sur le composant **tFileInputDelimited** dans l'espace de modélisation graphique. Lorsqu'une fenêtre vous propose de propager les modifications au composant de sortie, cliquez sur **Yes**.

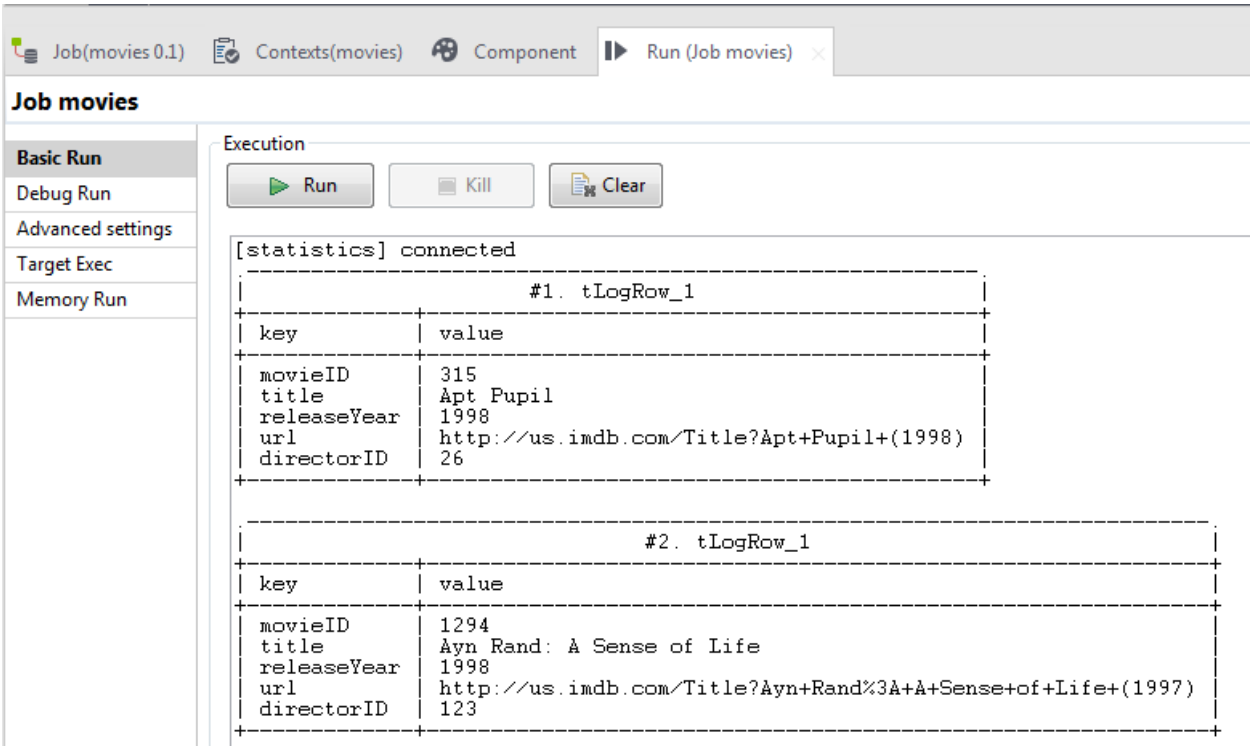
   Dans l'onglet **Basic settings** de la vue **Component**, vous pouvez voir que tous les paramètres du composant ont été automatiquement renseignés.



3. Double-cliquez sur le **tLogRow** pour ouvrir sa vue **Basic settings**.
4. Dans la zone **Mode**, sélectionnez l'option **Vertical (each row is a key/value list)** pour une meilleure lisibilité dans la console de la vue **Run**.



5. Appuyez sur **F6** ou cliquez sur le bouton **Run** de la vue **Run** pour exécuter votre Job.



Résultats

La console de la vue **Run** affiche les informations relatives aux films lues depuis le fichier source.

Filtrer les informations relatives aux films

Ce scénario étend le Job décrit dans Lire les informations relatives à des films à partir d'un fichier CSV à la page 12 afin de filtrer le flux de données et obtenir uniquement les données des films ayant des informations valides relatives aux réalisateurs.

This scenario demonstrates:

- Dupliquer un Job. Consultez Dupliquer le Job existant à la page 25 pour plus de détails.

- Ajouter un composant en saisissant son nom sur un lien ou dans l'espace de modélisation graphique . Consultez Ajouter un composant de mapping à la page 26 pour plus de détails.
- Déposer une métadonnée ou son schéma en tant que composant dans l'espace de modélisation graphique . Consultez Ajouter un composant lookup à la page 28 pour plus de détails.
- Effectuer un traitement simple sur des flux de données à l'aide du **tMap**. Consultez Configurer le mapping et exécuter le Job à la page 30 pour plus de détails.

Préparer la métadonnée relative aux réalisateurs

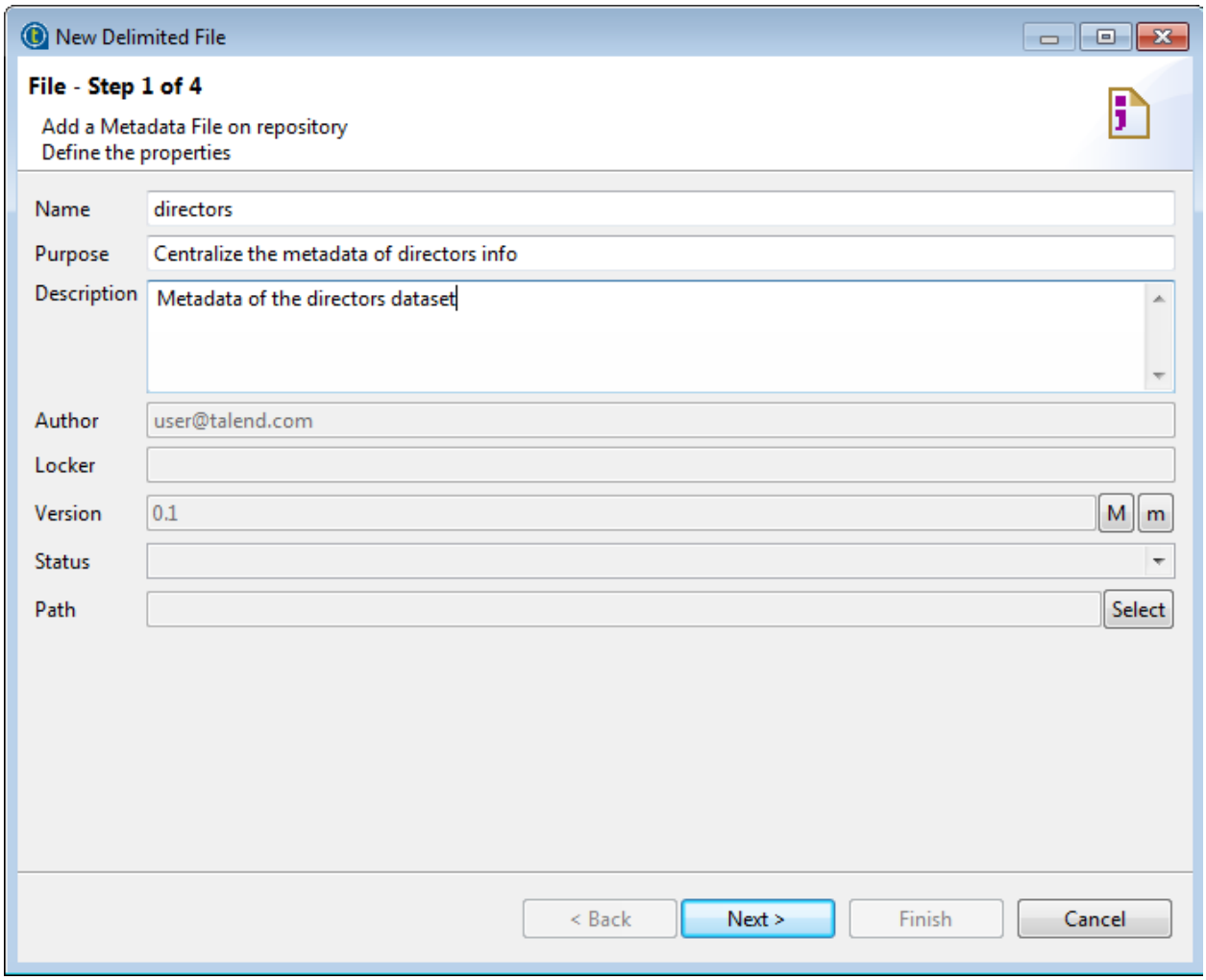
Cette procédure présente comment configurer la métadonnée du fichier de référence *directors.txt*directors.txt dans le **Repository**. Cette métadonnée sera utilisée pour ajouter et configurer l'entrée de référence dans le scénario.

Avant de commencer

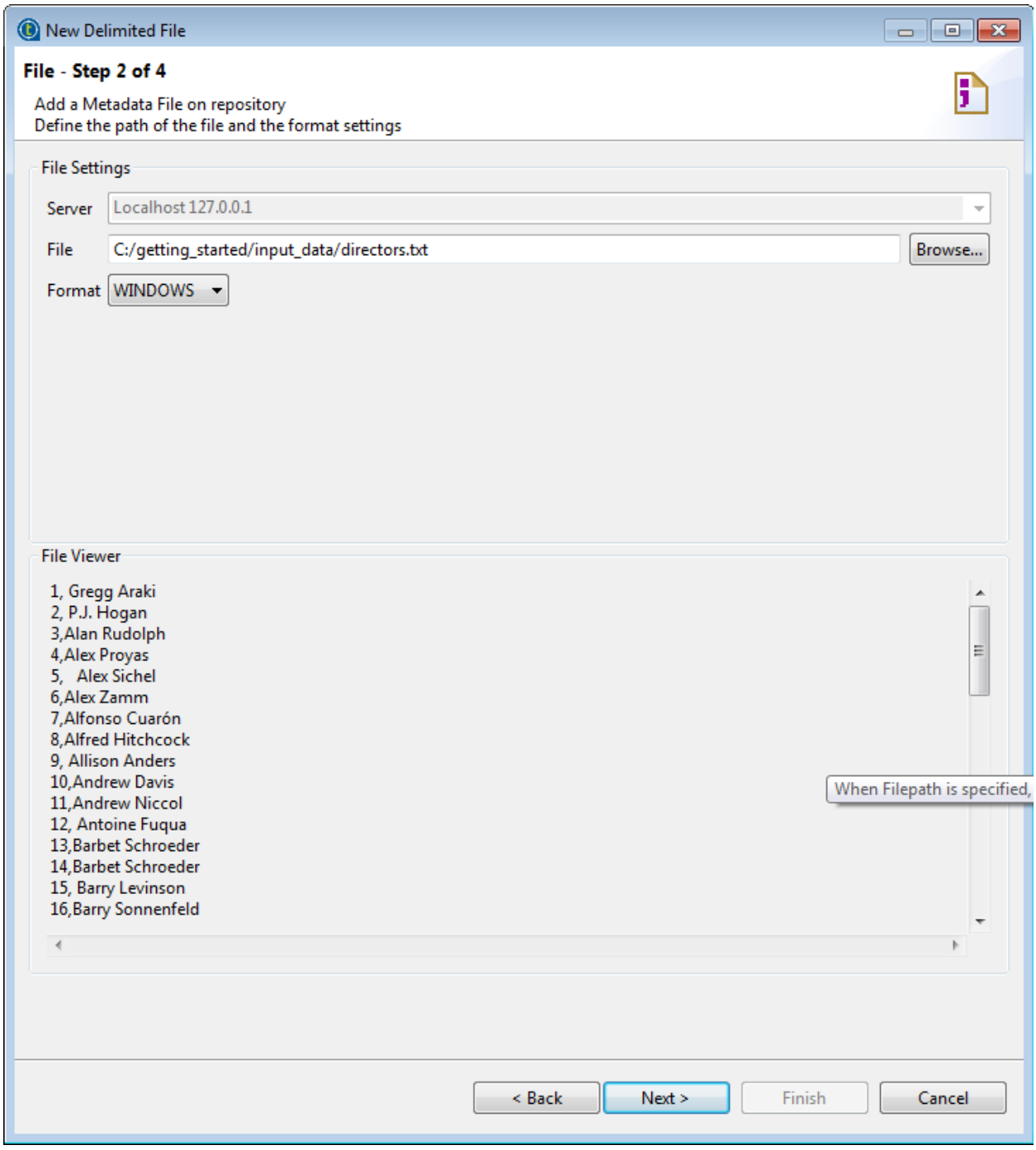
- Votre fichier `directors.txt` doit être disponible dans le dossier `C:\getting_started\input_data\`.

Procédure

1. Dans la vue **Repository**, développez le nœud **Metadata**, cliquez-droit sur **File delimited** et sélectionnez **Create file delimited** dans le menu contextuel pour ouvrir l'assistant **New Delimited File**.
2. Dans l'assistant **New Delimited File**, saisissez un nom pour la métadonnée, `directors` dans cet exemple et toutes les informations utiles permettant de décrire votre métadonnée. Cliquez sur **Next** pour passer à l'étape suivante et définir les propriétés générales du fichier.

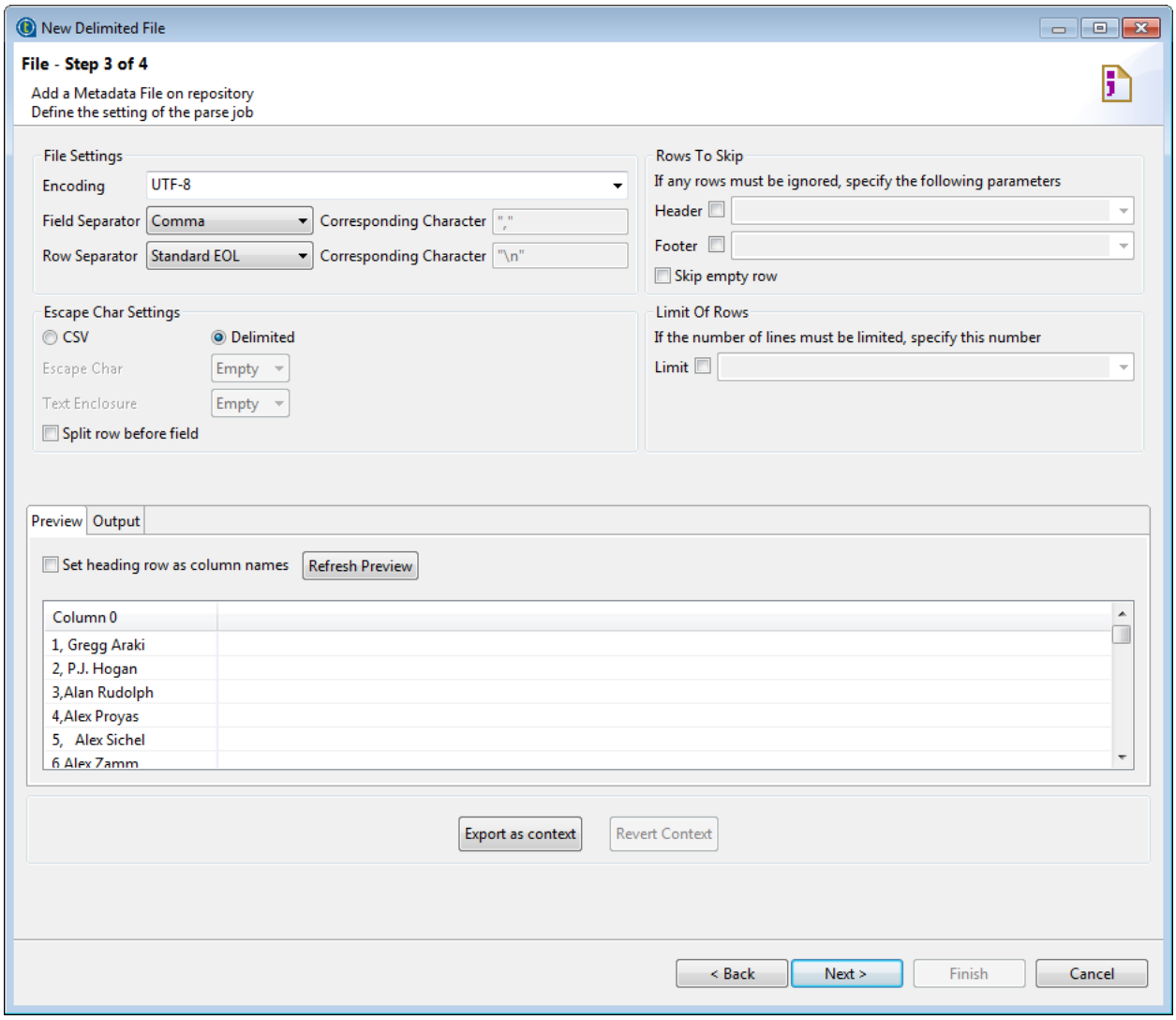


3. Dans le champ **File**, spécifiez le chemin du fichier source, ou cliquez sur **Browse** pour parcourir votre système jusqu'à ce fichier.



La zone **File Viewer** affiche un aperçu du fichier, vous permettant de vérifier sa cohérence, la présence d'un en-tête et la structure du fichier.

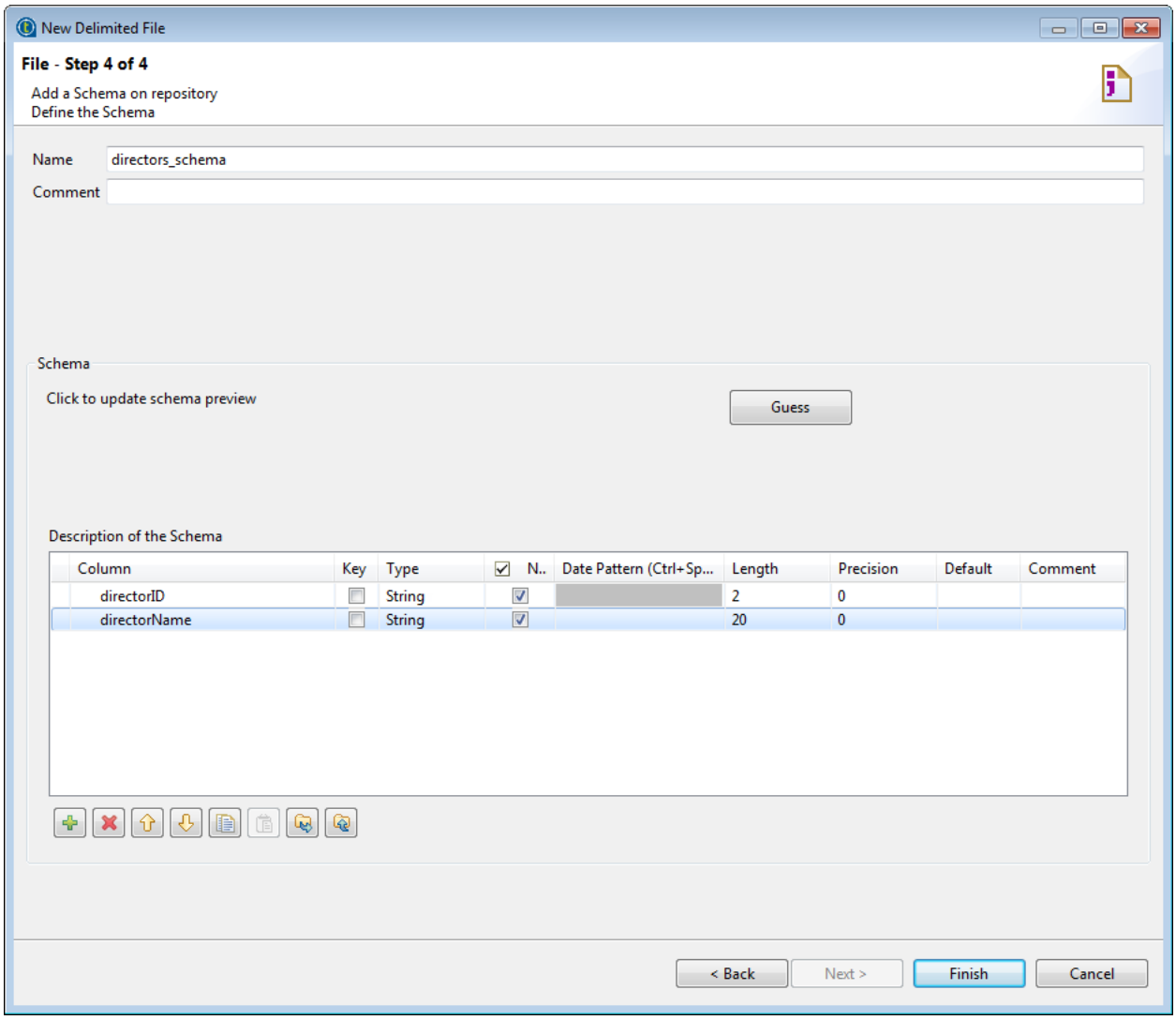
4. Sélectionnez **Windows** dans la liste **Format**, puis cliquez sur **Next** pour parser le fichier.
5. Dans la liste **Field Separator** de la zone **File Settings**, sélectionnez `Comma`.



6. Cliquez sur **Next** pour récupérer le schéma du fichier.

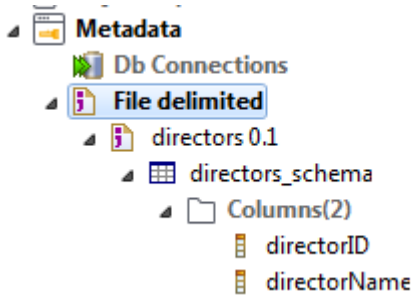
   La table **Description of the Schema** affiche le schéma généré du fichier.

7. Nommez le schéma `directors_schema`, et renommez les colonnes en `directorID` et `directorName`, respectivement, puis modifiez le type de données de la colonne `directorID` d'Integer à String.



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

   La métadonnée créée s'affiche dans la vue **Repository**.



Résultats

La métadonnée du fichier relatif aux réalisateurs est prête à être utilisée lors de la configuration des composants, pour lire ce fichier de référence.

Dupliquer le Job existant

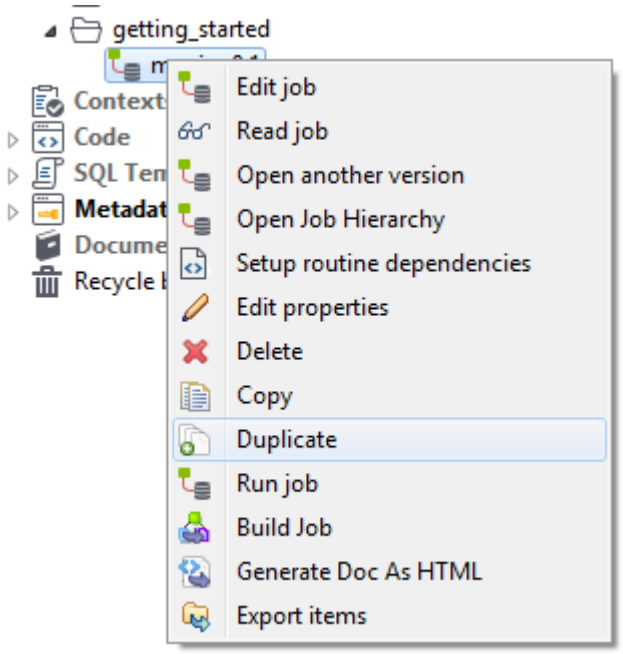
Cette procédure vous présente comment créer un Job à partir d'un Job existant.

Avant de commencer

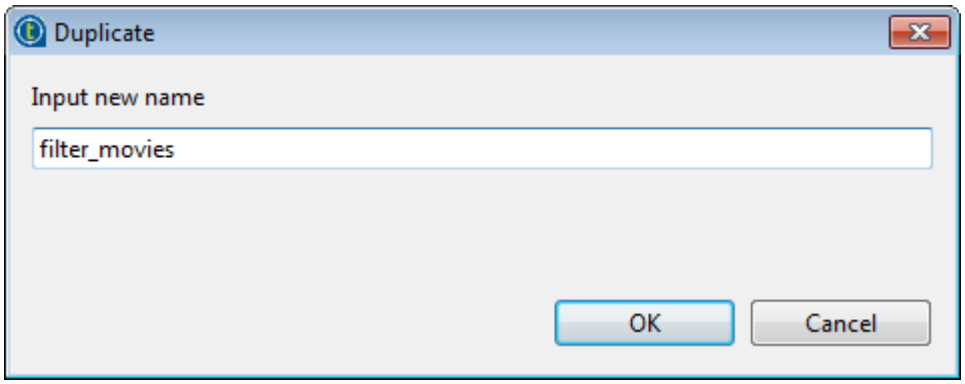
- Vous devez avoir créé et exécuté avec succès le Job nommé movies comme décrit dans Lire les informations relatives à des films à partir d'un fichier CSV à la page 12.

Procédure

1. Dans la vue **Repository**, cliquez-droit sur le Job nommé movies et sélectionnez **Duplicate** dans le menu contextuel.



2. Dans la boîte de dialogue **Duplicate**, saisissez un nom pour le Job, `filter_movies` dans cet exemple et cliquez sur **OK** pour valider la création du Job et fermer la boîte de dialogue.



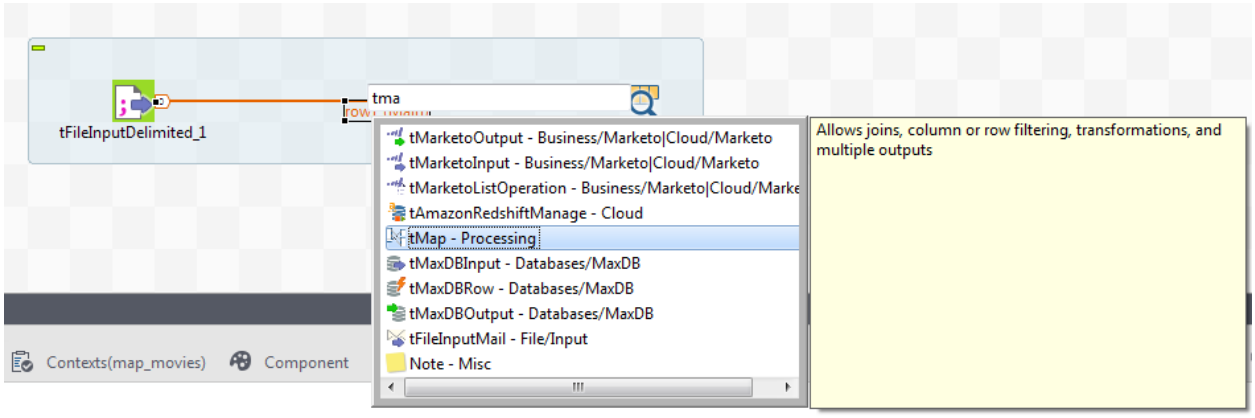
Le Job nommé filter_movies est créé, c'est un doublon du Job nommé movies.

Ajouter un composant de mapping

La procédure ci-dessous présente comment ajouter un composant de mapping en saisissant le nom du composant directement sur le lien existant.

Procédure

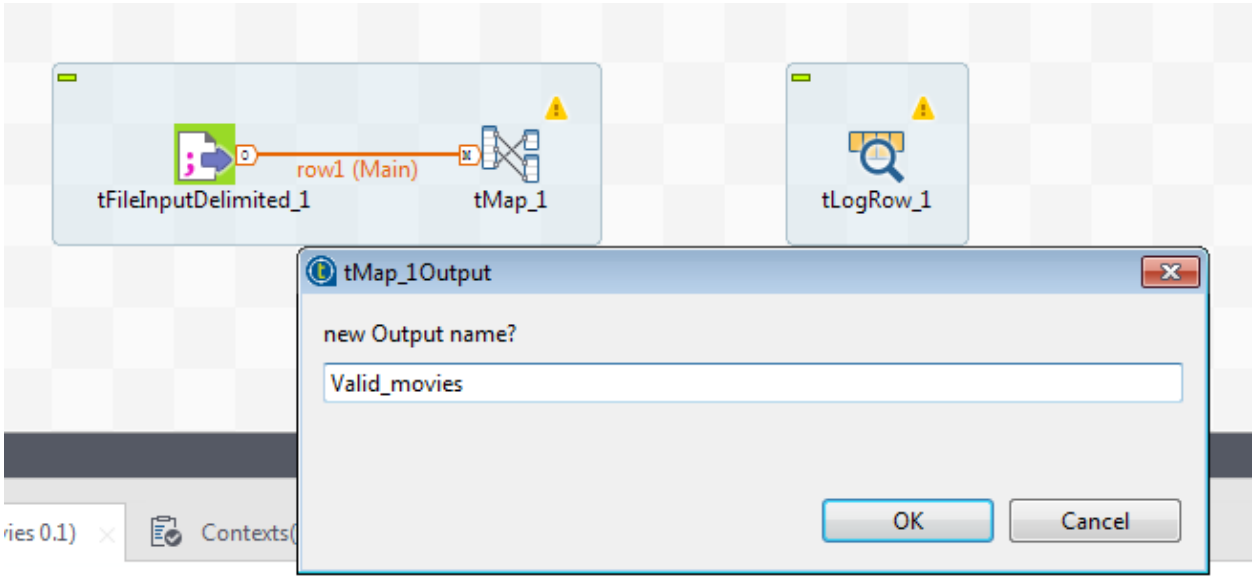
1. Dans le nouveau Job filter_movies, sélectionnez le lien **Row** reliant le **tFileInputDelimited** et le **tLogRow** et saisissez le nom du composant **tMap** ou une partie de son nom.



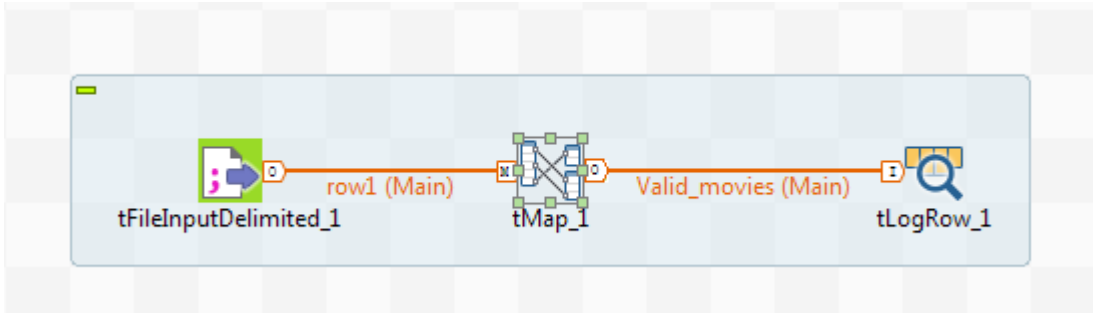
Lorsque vous commencez à saisir un nom de composant, une liste de composants correspondant à votre recherche s'affiche. Vous pouvez en sélectionner un pour voir sa description, à côté de la liste.

2. Double-cliquez sur le **tMap** dans la liste pour l'ajouter sur le lien.

   Le nouveau composant **tMap** est relié au composant d'entrée. Une boîte de dialogue s'ouvre et vous demande de saisir un nom pour le nouveau lien de sortie.



3. Saisissez un nom pour ce lien de sortie, `Valid_movies` dans cet exemple, puis cliquez sur **OK**. Lorsqu'il vous est proposé de propager le schéma d'entrée au composant cible de sortie, cliquez sur **Yes**.



Résultats

Le **tMap** est ajouté au Job et relié aux deux composants existants à l'aide d'un lien **Row** > **Main**.

Ajouter un composant lookup

La procédure ci-dessous vous explique comment ajouter un composant lookup (de référence) depuis le référentiel **Repository**, le relier au **tMap** et activer l'option supprimant les espaces.

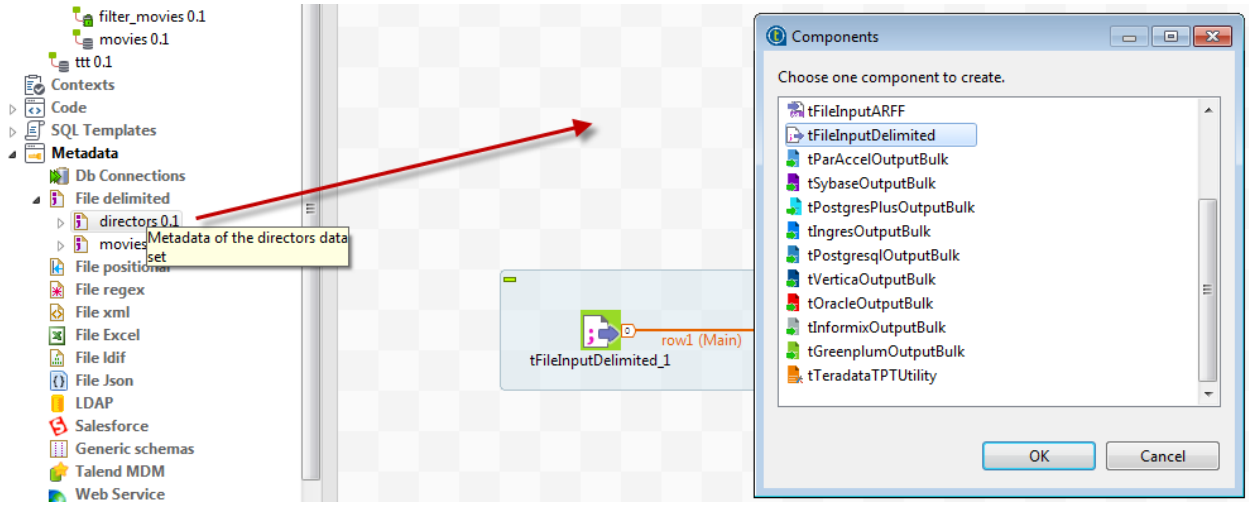
Avant de commencer

- Vous devez avoir centralisé la métadonnée du fichier `directors.txt` dans le **Repository**, comme décrit dans Préparer la métadonnée relative aux réalisateurs à la page 21.

Procédure

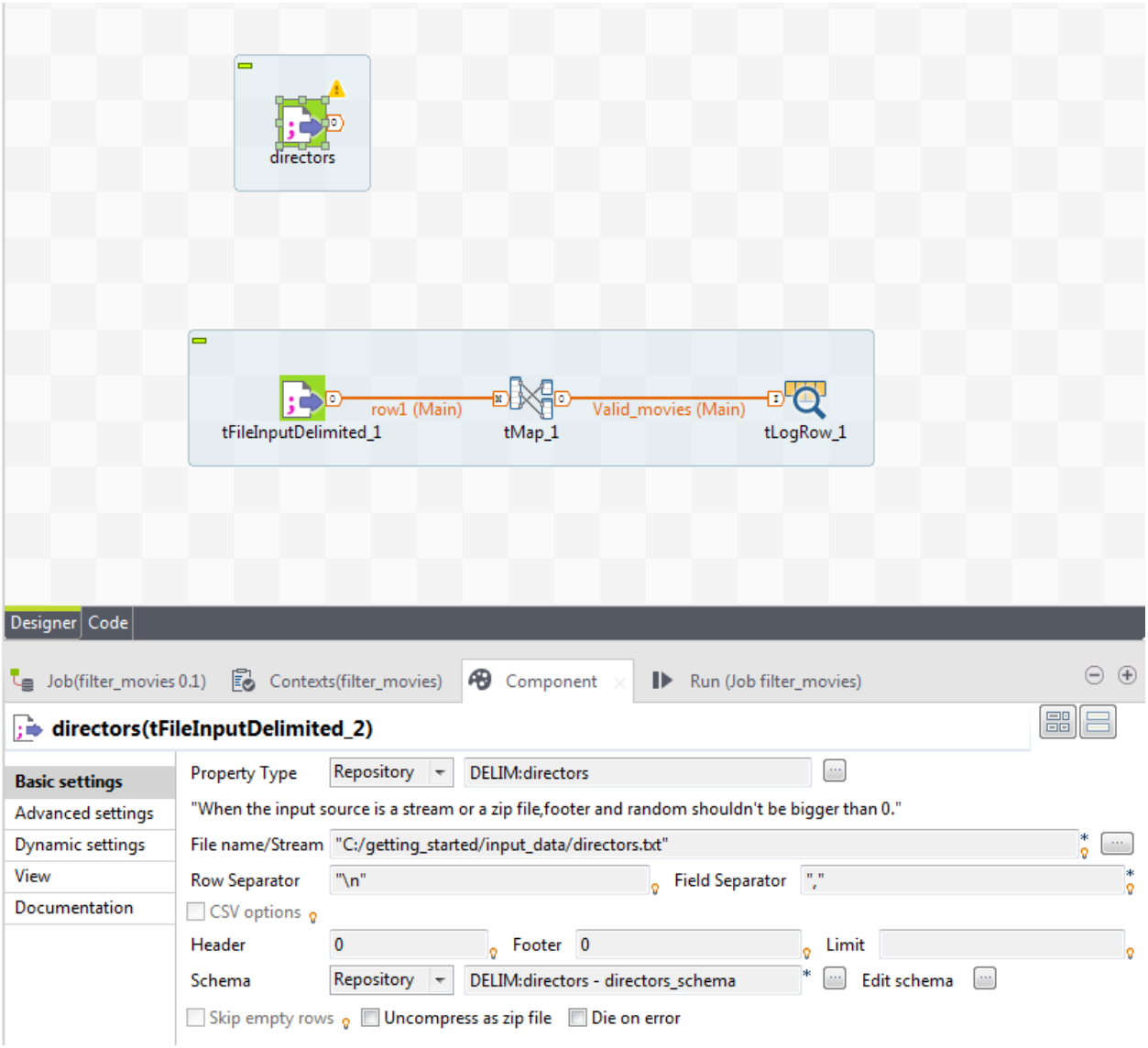
1. Dans la vue **Repository**, développez **Metadata** > **File delimited**, glissez-déposez la métadonnée **directors** ou son schéma **directors_schema** dans l'espace de modélisation graphique.

   La boîte de dialogue **Components** s'ouvre, affichant une liste de composants que vous pouvez ajouter au Job à partir de cette métadonnée.



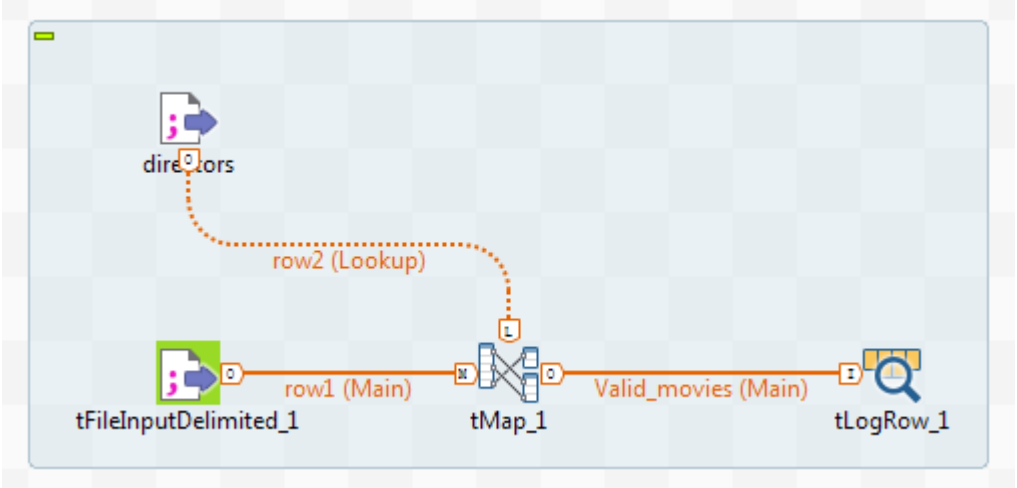
2. Sélectionnez le **tFileInputDelimited** et cliquez sur **OK**.

   Un **tFileInputDelimited** nommé **directors** est ajouté à l'espace de modélisation graphique et ses paramètres simples (onglet **Basic settings**) sont automatiquement renseignés.



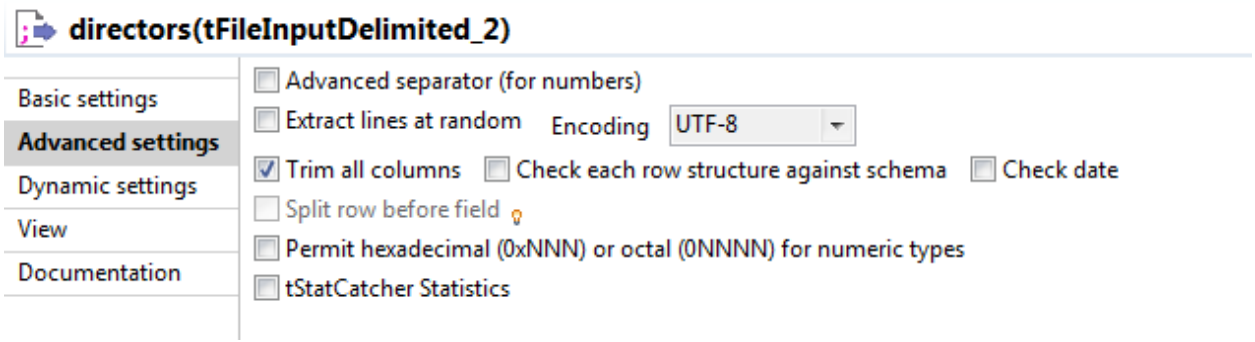
3. Cliquez-droit sur le nouveau **tFileInputDelimited**, sélectionnez **Row** > **Main** dans le menu contextuel et cliquez sur le **tMap**.

   Le **tFileInputDelimited** est relié au **tMap** à l'aide d'un lien **Lookup**.



4. Dans l'onglet **Advanced settings** du nouveau **tFileInputDelimited** et cochez la case **Trim all columns**.

Certains enregistrements du fichier d'entrée de référence `directors.txt` contiennent des espaces blancs en début de champ. Cette option vous permet de supprimer ces espaces blancs du flux de référence lorsque le Job est exécuté.



Résultats

Votre Job contient tous les composants nécessaires pour filtrer les informations relatives aux films. Vous allez ensuite configurer le mapping dans le composant **tMap** afin de filtrer le flux d'entrée principal par rapport au flux de référence et écrire en sortie les informations souhaitées.

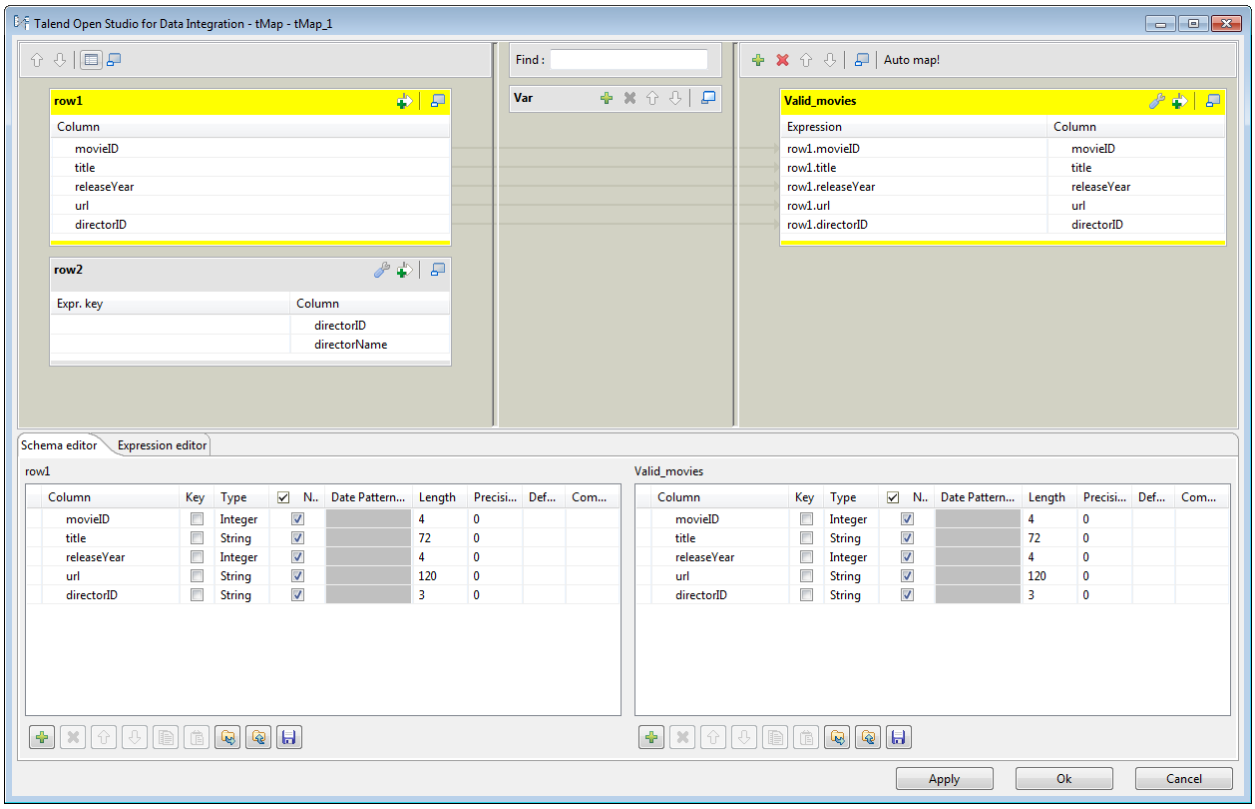
Configurer le mapping et exécuter le Job

La procédure ci-dessous vous apprend à configurer les mappings et les jointures Inner Join pour écrire en sortie les informations relatives aux films ayant un ID de réalisateur valide.

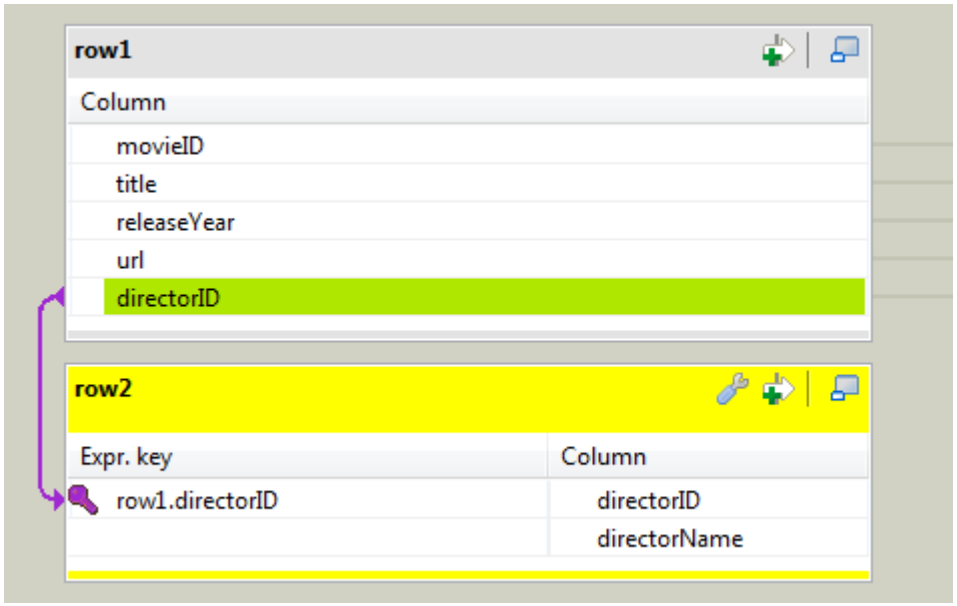
Procédure

1. Double-cliquez sur le composant **tMap** pour ouvrir son éditeur de mapping.

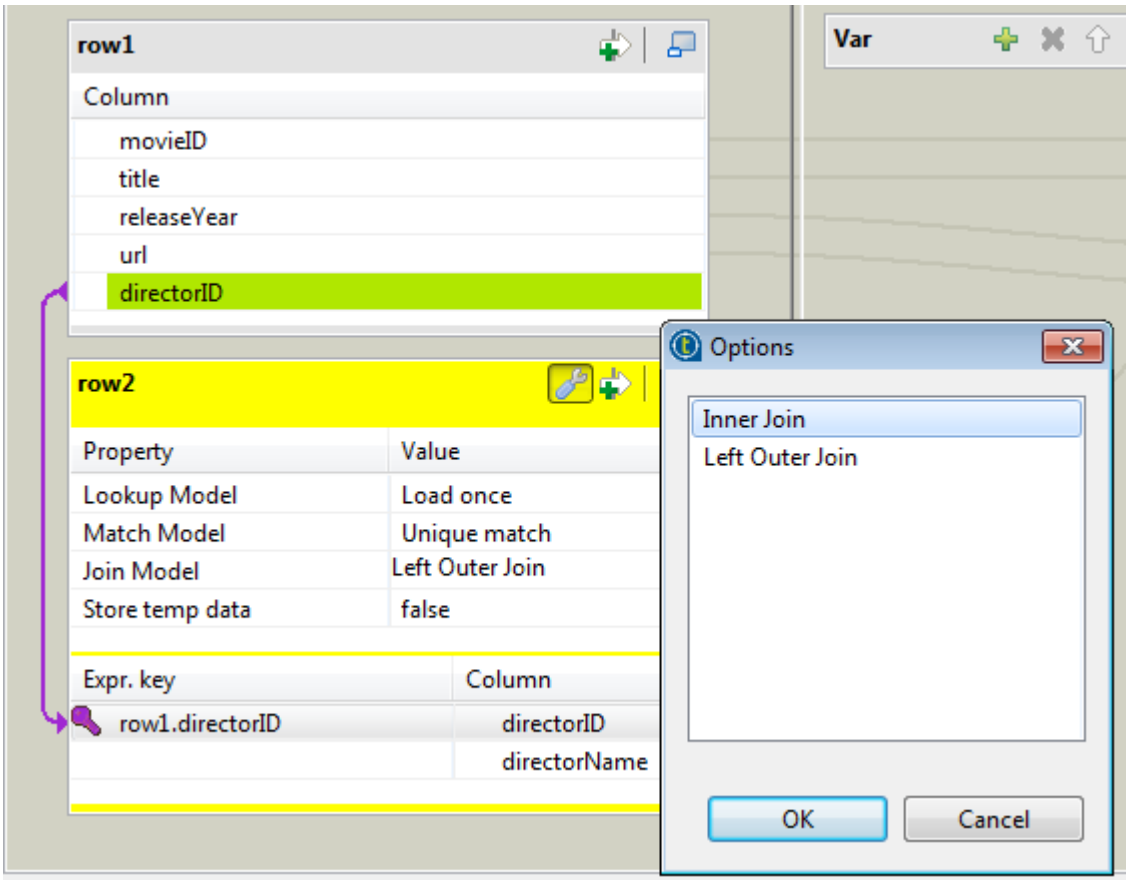
   L'éditeur de mapping affiche trois tables, nommées **row1**, **row2** et **Valid_movies** dans cet exemple, correspondant respectivement au schéma du fichier des films, au schéma du fichier des réalisateurs et au schéma de sortie des informations valides. Les colonnes de la table *row1*row1 sont déjà mappées aux colonnes de la table *Valid_movies*Valid_movies.



2. Sélectionnez la colonne **directorID** de la table **row1** et glissez-la sur la colonne **directorID** dans la table **row2** afin de créer une jointure entre les deux ensembles de données basée sur l'ID des réalisateurs.

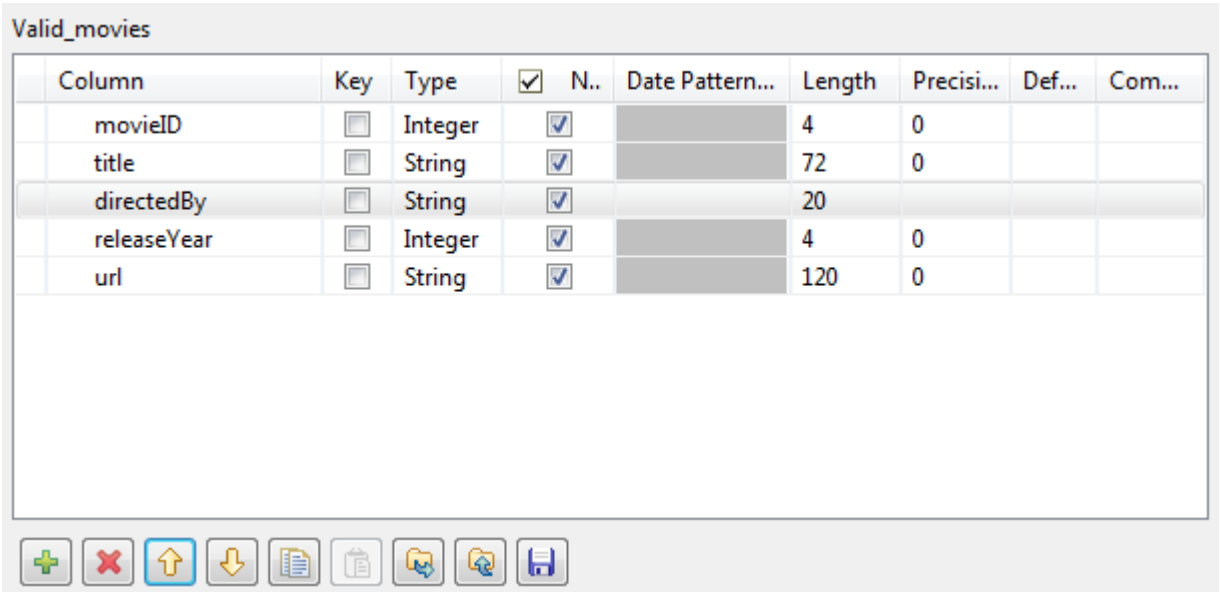


3. Cliquez sur le bouton **tMap settings** (représentant une clé anglaise), cliquez sur le champ **Value** pour **Join Model**, puis cliquez sur le bouton **[...]** qui s'affiche pour ouvrir la boîte de dialogue **Options**. Dans la boîte de dialogue, sélectionnez **Inner Join** et cliquez sur **OK** pour définir la jointure comme Inner Join.



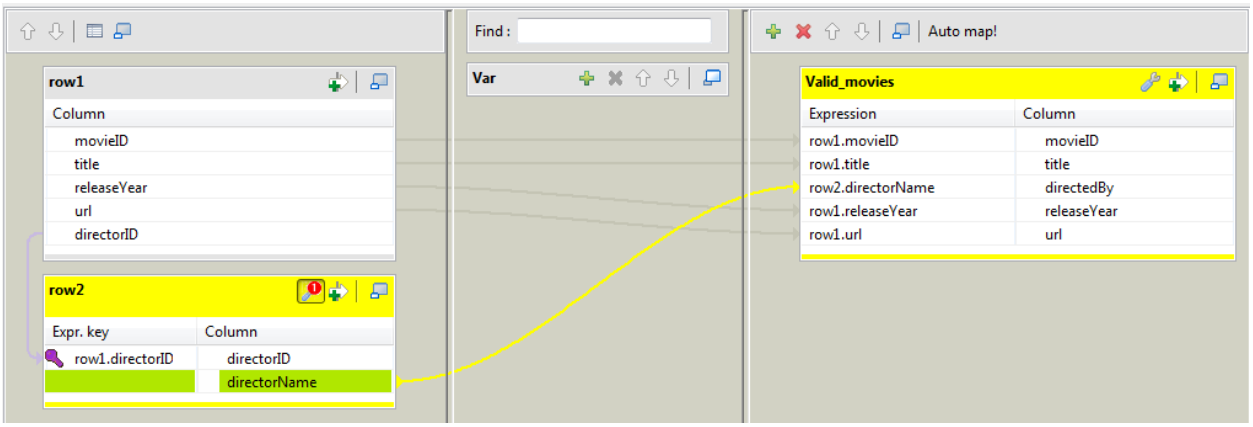
Grâce à ce paramètre, seuls les enregistrements de films dont l'ID du réalisateur correspond à ceux du fichier de référence seront passés au composant de sortie.

4. Dans la zone **Schema editor** au bas de l'éditeur de mapping, sélectionnez la colonne **directorID** du schéma de sortie, **Valid_movies** dans cet exemple et cliquez sur le bouton **[x]** afin de la supprimer.
5. Cliquez sur le bouton **[+]** sous la table de sortie pour ajouter une colonne, nommez-la `directedBy`, configurez sa longueur **Length** à `20`, puis déplacez-la pour la placer entre **title** et **releaseYear**.



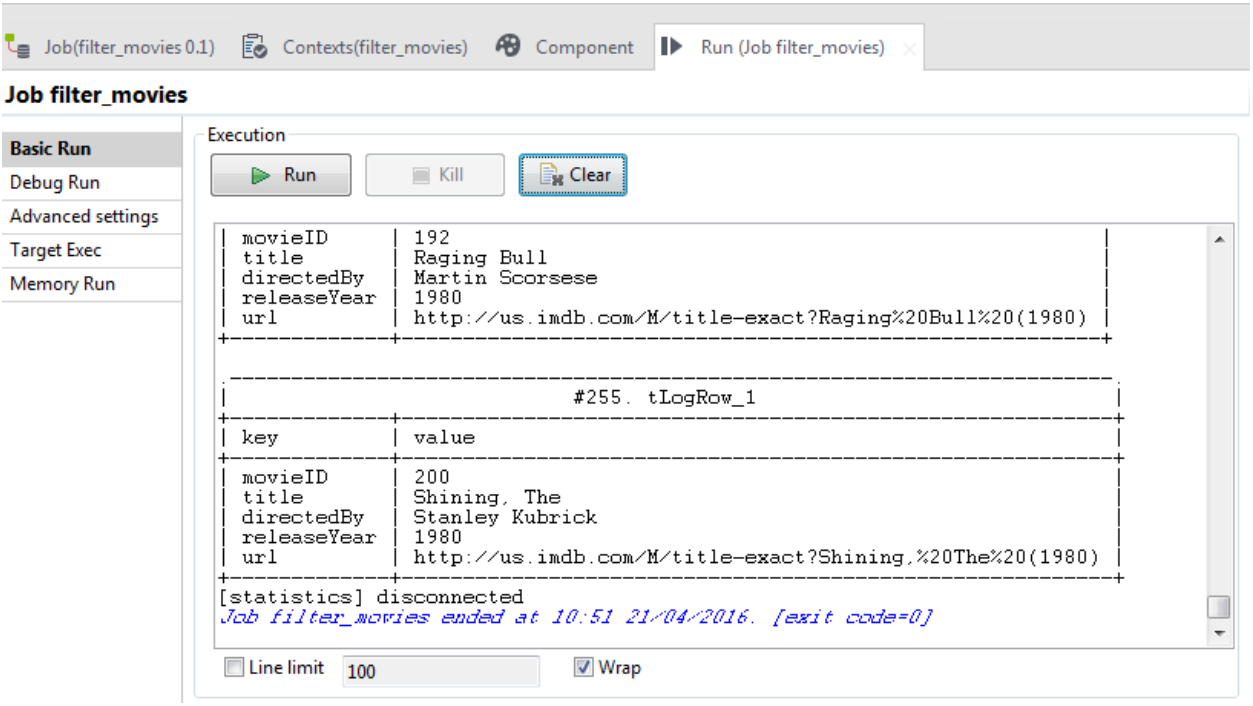
6. Sélectionnez la colonne **directorName** de la table **row2** et glissez-la dans le champ **Expression** correspondant à la colonne **directedBy** dans la table de sortie.

   Un nouveau mapping est créé entre la table de référence et la table de sortie.



7. Cliquez sur **OK** pour valider les mappings et fermer l'éditeur, puis cliquez sur **Yes** lorsqu'il vous est proposé de propager les modifications.

   La configuration des mappings est sauvegardée et le schéma de sortie est synchronisé au composant de sortie **tLogRow**.
8. Appuyez sur **F6** ou cliquez sur le bouton **Run** de la vue **Run** pour exécuter le Job.



Résultats

Seuls les enregistrements de films ayant des informations valides relatives aux réalisateurs sont affichées dans la console de la vue **Run**.

Collecter les informations rejetées relatives à des films et sauvegarder les résultats de traitement dans une base de données

À partir du scénario décrit dans Filtrer les informations relatives aux films à la page 20, ce scénario agrandit le Job afin de rassembler les données des films dans lesquelles manquent les informations des réalisateurs et afin d'écrire les données valides et invalides dans une base de données MySQL.

Ce scénario montre :

- Comment ajouter un composant en saisissant son nom dans l'espace de modélisation graphique ou en le glissant depuis un composant existant. Consultez Ajouter des composants de sortie de base de données à votre Job à la page 34 pour plus de détails.
- Comment configurer les mappings pour les informations rejetées dans le **tMap**. Consultez Configurer le mapping pour les données rejetées à la page 36 pour plus de détails.
- Comment configurer les sorties de base de données. Consultez Configurer les sorties de base de données MySQL à la page 38 pour plus de détails.

Ajouter des composants de sortie de base de données à votre Job

Dans l'exemple ci-dessous, vous allez créer un nouveau Job à partir du Job **filter_movies** et ajouter deux composants **tMysqlOutput**. Ces composants seront utilisés pour écrire les informations des films traitées dans les tables de base de données spécifiées.

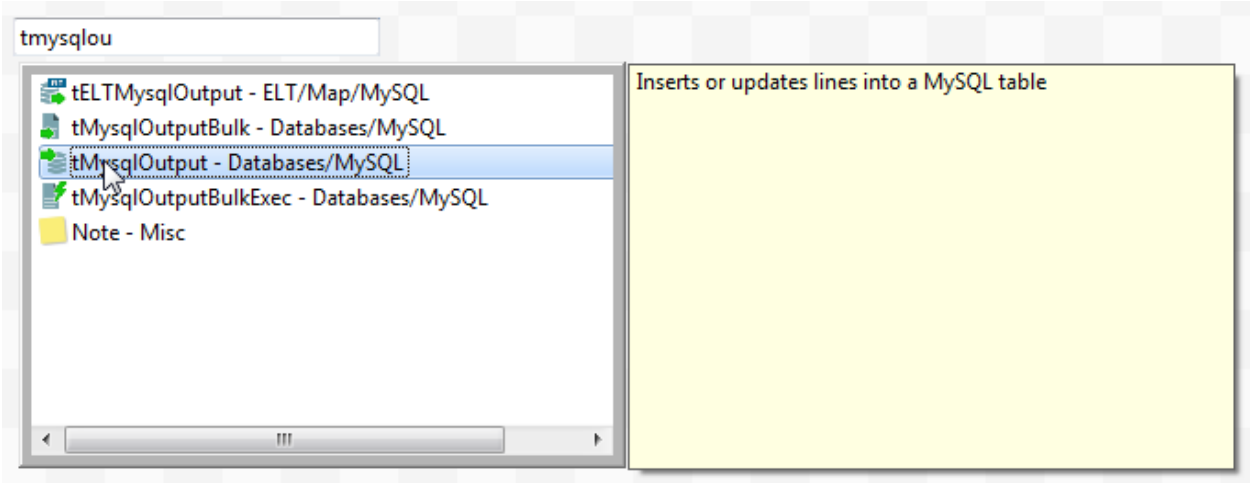
Avant de commencer

- Vous devez avoir créé et exécuté avec succès le Job **filter_movies** comme décrit dans Filtrer les informations relatives aux films à la page 20.

Procédure

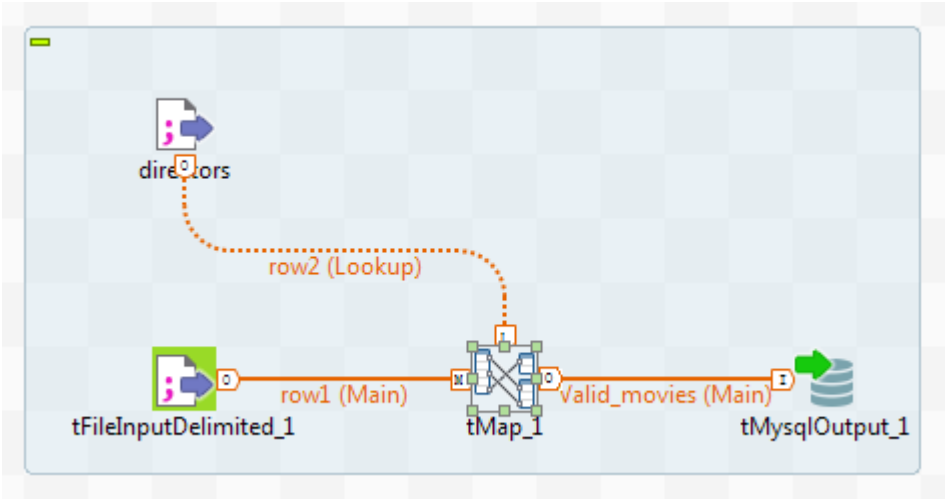
1. Créez un nouveau Job en dupliquant le Job créé dans le scénario précédent et nommez le nouveau Job `write_movies_to_db`, puis double-cliquez sur le Job pour l'ouvrir dans l'espace de modélisation graphique.
2. Cliquez-droit sur le composant **tLogRow** et sélectionnez **Delete** dans le menu contextuel pour le supprimer.
3. Cliquez à l'ancien emplacement du **tLogRow** dans l'espace de modélisation graphique et saisissez le nom du **tMysqlOutput** ou une partie de celui-ci puis, sélectionnez et double-cliquez sur le **tMysqlOutput** dans la liste pour l'ajouter dans l'espace de modélisation graphique.

   Lorsque vous commencez à saisir un nom de composant, une liste de composants correspondant à votre recherche s'affiche. Vous pouvez en sélectionner un pour voir sa description, à côté de la liste.



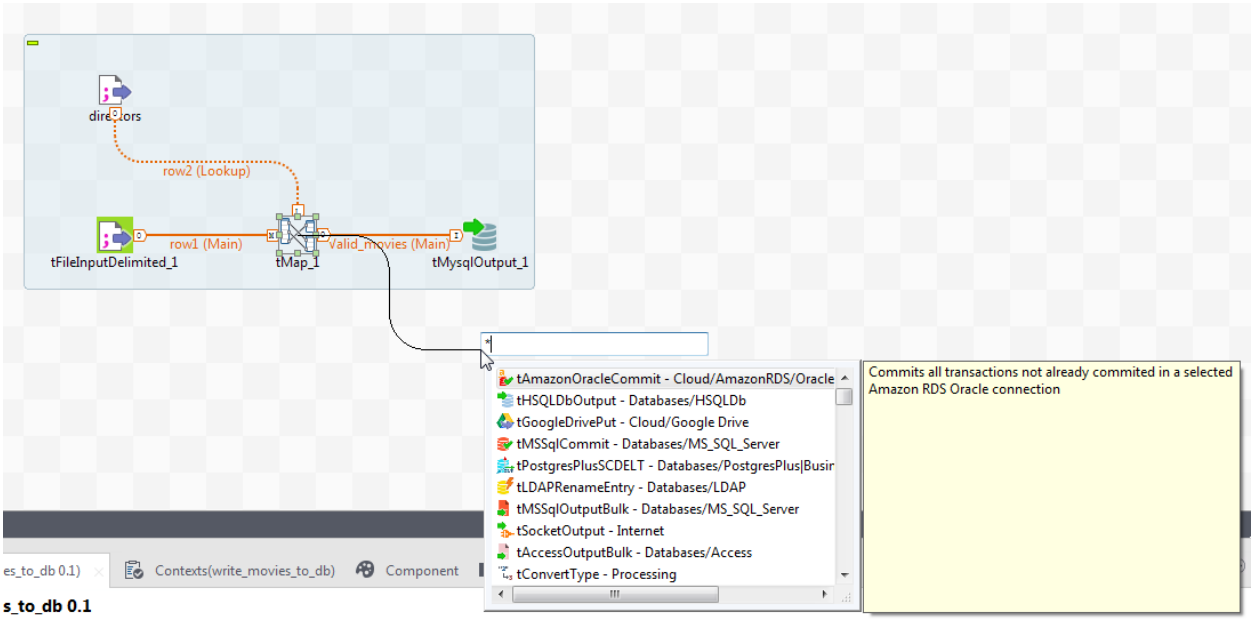
4. Cliquez-droit sur le composant **tMap**, sélectionnez **Row** > **Valid_movies** dans le menu contextuel et cliquez sur le **tMysqlOutput** pour le relier au **tMap**.

Le nom de la connexion **Valid_movies** correspond au nom de la table de sortie existante dans le **tMap**.



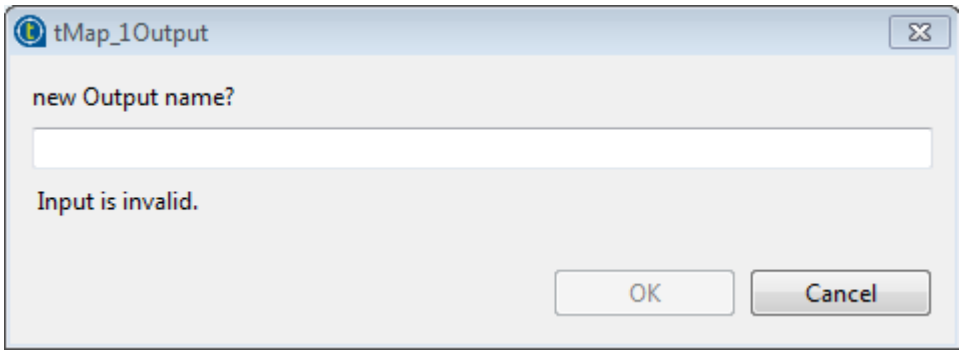
5. Cliquez sur le composant **tMap** et glissez-déposez l'icône **o** sur l'espace de modélisation graphique.

   Un champ textuel et une liste de composants suggérés s'affichent. Vous pouvez en sélectionner un pour voir sa description, à côté de la liste.

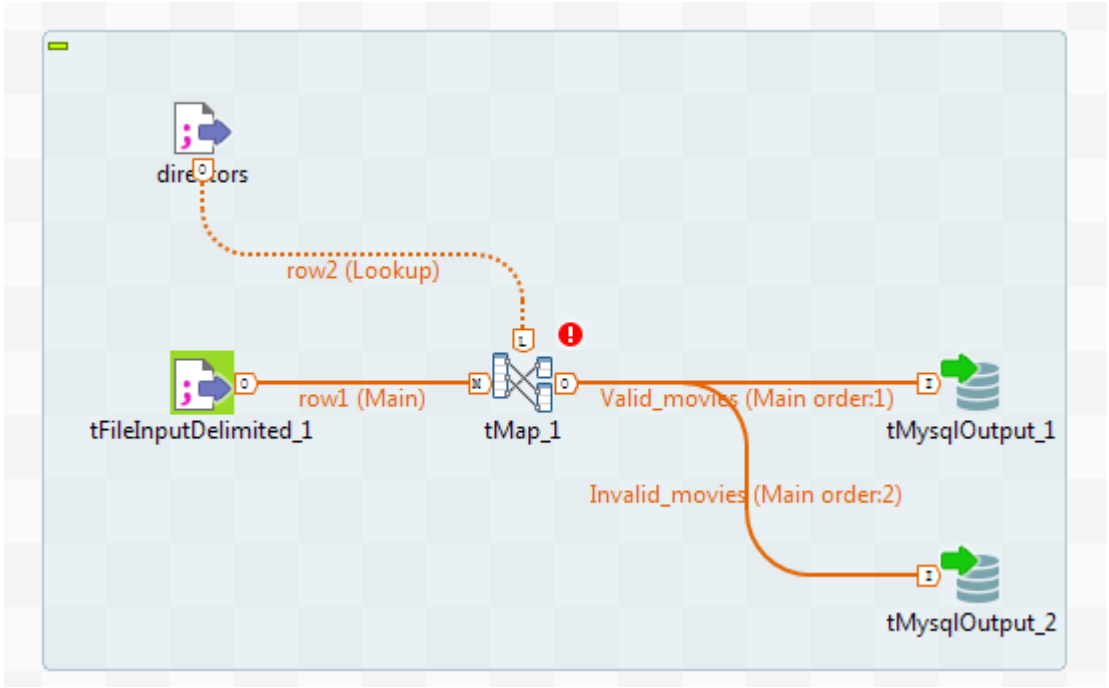


6. Dans le champ textuel, saisissez le nom du **tMysqlOutput**, sélectionnez le composant dans la liste et appuyez sur **Entrée** pour ajouter un autre composant **tMysqlOutput** dans l'espace de modélisation graphique.

   Une boîte de dialogue s'ouvre et vous demande de saisir un nom pour la connexion de sortie.



7. Dans la boîte de dialogue, saisissez `Invalid_movies` et cliquez sur **OK** pour relier le **tMap** au deuxième composant **tMysqlOutput**.



Résultats

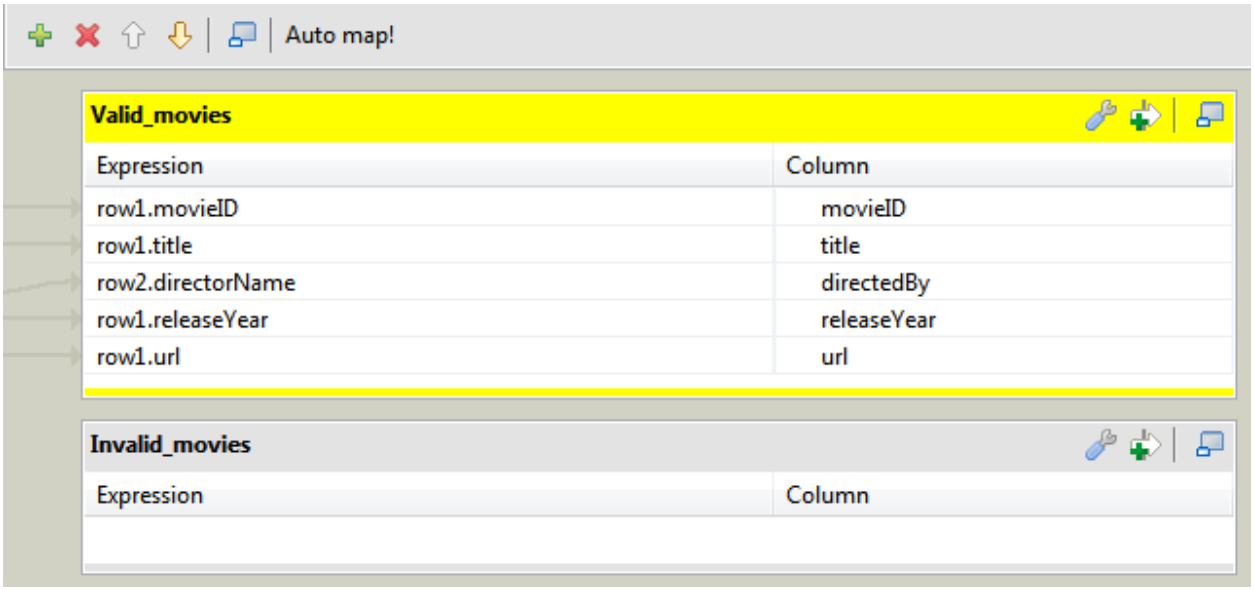
Vous avez ajouté et connecté les composants de sortie de base de données nécessaires pour écrire les informations des films traitées dans une base de données MySQL. Maintenant, vous devez configurer de nouveaux mappings dans le **tMap** et les paramètres de base de données dans les composants **tMysqlOutput**.

Configurer le mapping pour les données rejetées

La procédure ci-dessous vous explique comment configurer les mappings pour rassembler les informations rejetées.

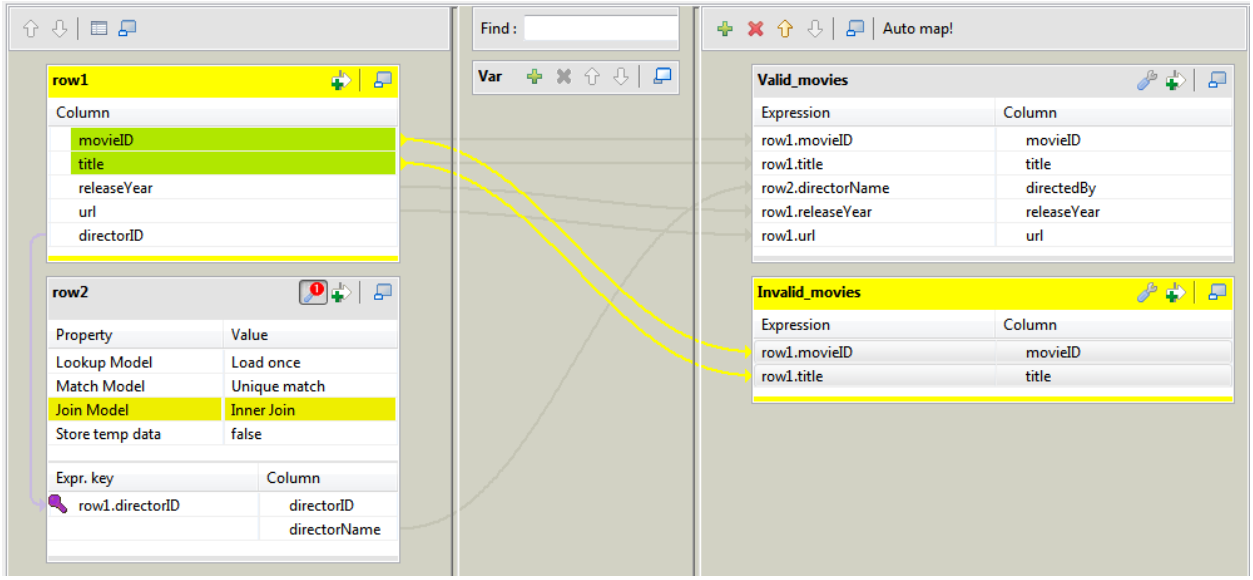
Procédure

1. Double-cliquez sur le composant **tMap** pour ouvrir l'éditeur **Map Editor**.

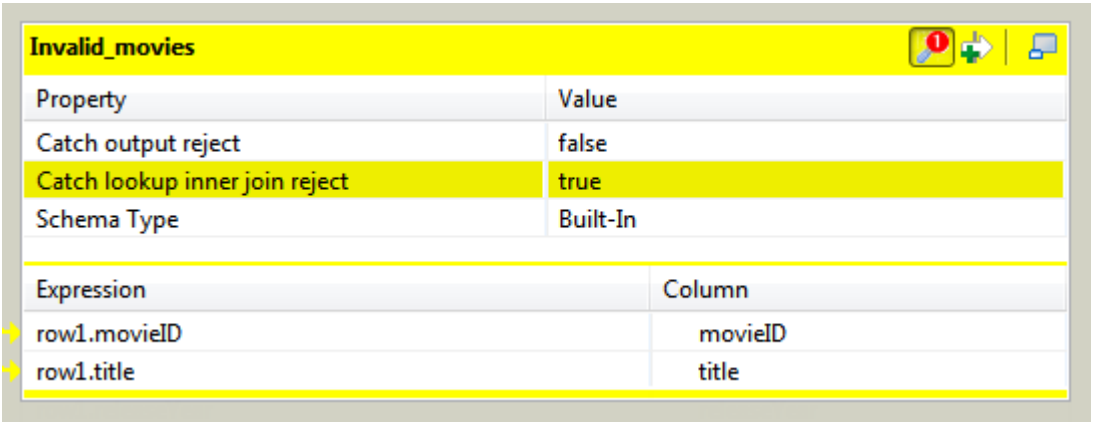


Une deuxième table de sortie nommée **Invalid_movies** a été automatiquement créée.

2. Déposez les colonnes **movieID** et **title** à partir de la table **row1** vers la table **Invalid_movies**.



3. Cliquez sur le bouton **tMap settings** de la table **Invalid_movies** puis sur le champ **Value** pour **Catch lookup inner join reject** et cliquez sur le bouton **[...]** qui s'affiche pour ouvrir la boîte de dialogue **Options**. Dans la boîte de dialogue, sélectionnez **true** et cliquez sur **OK**.



Grâce à ce paramètre, les enregistrements sans l'ID du réalisateur ou avec un ID ne correspondant pas à ceux du fichier de référence seront passés au composant de sortie.

4. Cliquez sur **OK** pour valider les mappings et fermez l'éditeur **Map Editor** puis, cliquez sur **Yes** lorsqu'il vous est proposé de propager les modifications.

   La configuration des mappings est sauvegardée. Le schéma de sortie et le composant de sortie sont synchronisés.

Résultats

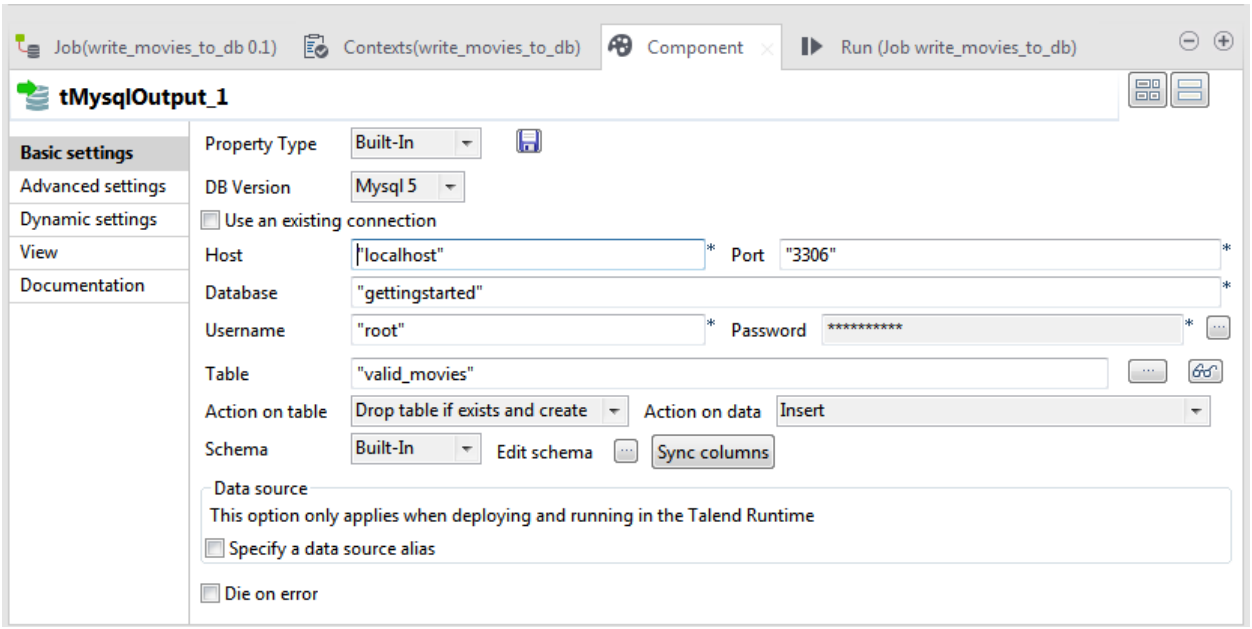
Vous avez configuré les mappings pour les sorties rejetées. Maintenant, vous devez configurer les composants de sortie pour écrire les flux de sortie des tables de base de données.

Configurer les sorties de base de données MySQL

La procédure ci-dessous vous explique comment configurer les composants de sortie de base de données pour écrire les informations des films des tables de base de données MySQL.

Procédure

1. Double-cliquez sur le premier composant **tMysqlOutput** pour ouvrir sa vue **Component**.



2. Fournissez les détails de connexion nécessaires pour accéder à votre base de données, à savoir le nom d'hôte ou l'adresse IP, le numéro de port, le nom de la base de données, le nom et le mot de passe de l'utilisateur dans les champs correspondants.

   Lorsque vous saisissez votre mot de passe, vous devez d'abord cliquer sur le bouton **[...]** puis sur le champ **Password** pour ouvrir une boîte de dialogue. Saisissez votre mot de passe entre guillemets doubles dans le champ textuel puis, cliquez sur **OK**.

3. Dans le champ **Table**, saisissez le nom de la table de base de données cible.

   Dans cet exemple, la table relative aux informations de films valides est `valid_movies`.

4. Dans les listes **Action on table** et **Action on data**, sélectionnez l'option répondant à vos besoins.

   Dans cet exemple, vous pouvez d'abord supprimer la table si elle existe déjà, en créer une nouvelle, vide et utilisez l'option par défaut de la liste **Action on data**.

5. Dans l'onglet **Basic settings** du deuxième composant **tMysqlOutput**, utilisez les mêmes paramètres que dans le premier **tMysqlOutput** sauf pour le nom de la table de base de données cible.

   Dans cet exemple, la table relative aux informations de films invalides est `invalid_movies`.
6. Appuyez sur **F6** ou cliquez sur le bouton **Run** de la vue **Run** pour exécuter votre Job.

Résultats

Les enregistrements de films dont l'ID du réalisateur est valide sont sauvegardés dans la table de base de données nommée valid_movies et ceux dont l'ID du réalisateur est invalide sont sauvegardés dans la table de base de données nommée invalid_movies.

Que faire ensuite ?

Vous avez vu comment le Studio Talend vous permet de gérer vos données à l'aide de Jobs Talend. Vous avez appris à accéder à vos données via le Studio Talend, les filtrer et les transformer, puis stocker les données filtrées et transformées dans une base de données. Vous avez également appris à centraliser les connexions fréquemment utilisées dans le **Repository** et les réutiliser facilement dans vos Jobs.

Pour en savoir plus au sujet du Studio Talend, consultez :

- le Guide utilisateur du Studio Talend
- la documentation des composants Talend

Pour vous assurer de la propreté de vos données, vous pouvez utiliser Talend Open Studio for Data Quality et Talend Data Preparation Free Desktop.

Pour en savoir plus au sujet des produits et solutions Talend, consultez fr.talend.com.