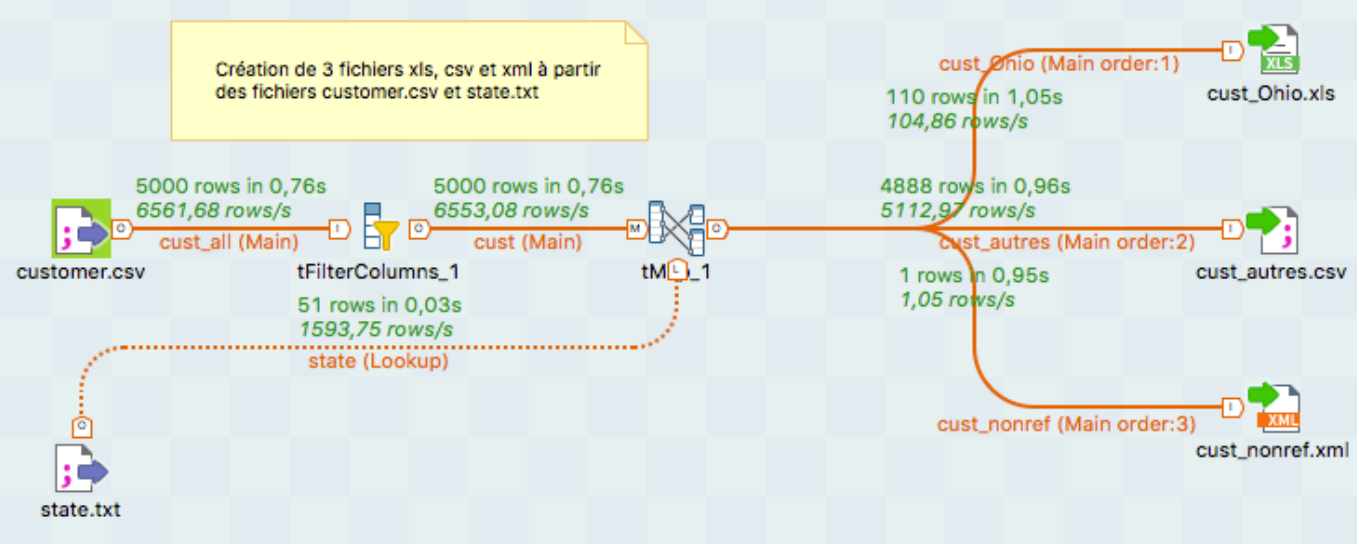
TP Talend Open Studio

Introduction à l'Intégration de Données

L'objectif de ce TP est de créer le job suivant :



Sources :

- customer.csv : liste de clients avec le numéro de l'état
- state.txt : numéro et nom des états américains

Cibles :

- Fichier cust_Ohio.xls : clients de l'état de l'Ohio
- Fichier cust_autres.csv : clients des autres états
- Fichier cust_nonref.xml : clients avec un état non référencé

Transformations :

- Ne garder que les colonnes id, CustomerName, CustomerAddress du fichier customer.csv et ajouter la colonne LabelState du fichier state.txt.
- Ne pas traiter les clients dont l'état est inconnu.

1) Création du projet :

- Ouvrir TOS_MDM-Studio.
- Créer un nouveau projet (dans la fenêtre suivante, vous pouvez ignorer la connexion à TalendForge).
- Choisir la perspective « Integration » (en haut à droite). Les perspectives Profiling et MDM seront utilisées pour la partie Gouvernance des Données.

La fenêtre de Talend Open Studio est composée des vues suivantes :

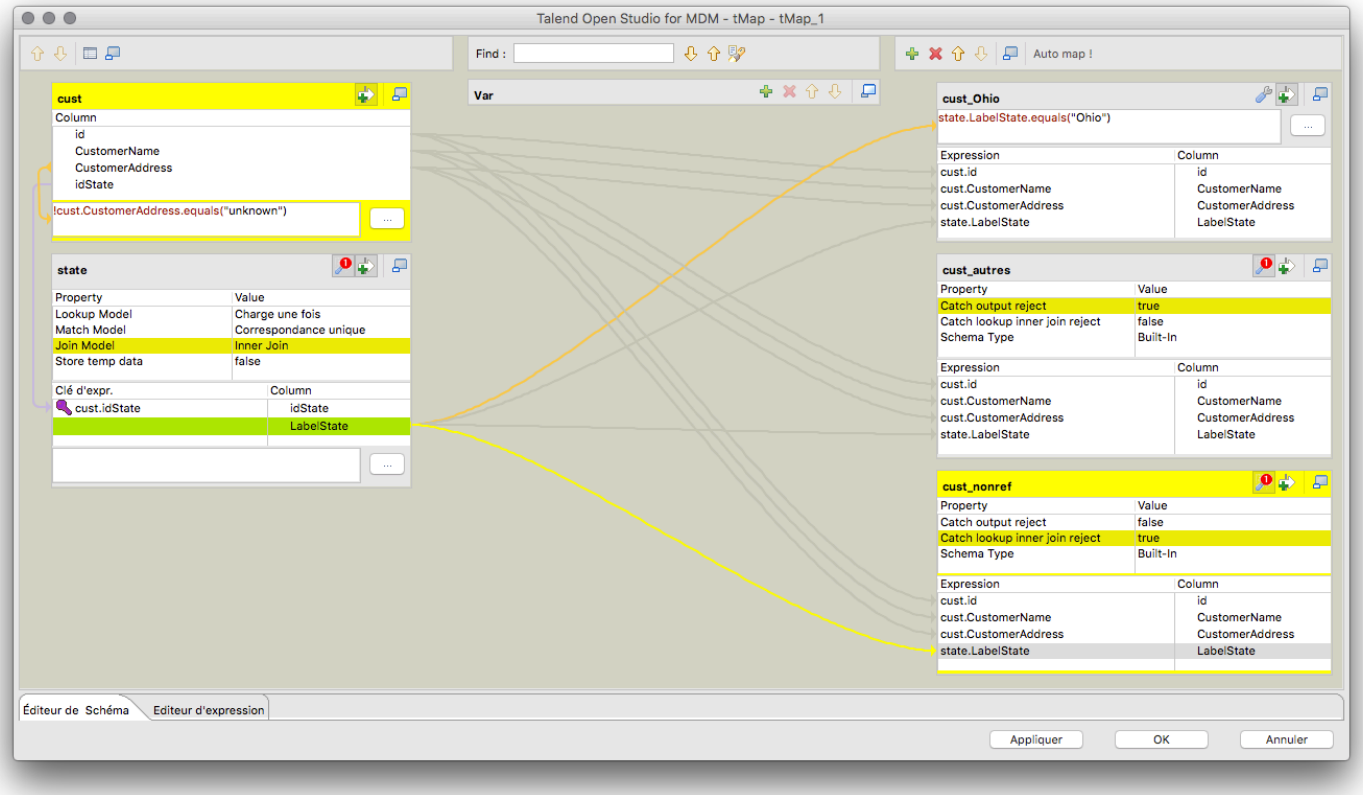
- Barres d'outils et menus (en haut).
- Repository/Référentiel (en haut à gauche) : Contient tous les éléments techniques du projet. C'est ici que seront définies les métadonnées.

- Designer (au centre) : Cet espace de modélisation permet de concevoir les jobs. L'onglet Code permet de voir le code java correspondant.
- Palette (à droite) : Cette palette graphique permet d'accéder aux différents composants.
- Différentes vues (en bas au centre) :
    - Job : infos sur le job sélectionné
    - Composant : configuration du composant sélectionné
    - Exécuter : exécution des jobs
- Outline et Aperçu (en bas à gauche) : Ces fenêtres fournissent un aperçu du code et du schéma du job.

2) Spécification des métadonnées des fichiers sources :
- Dans le Repository/Métadonnées, spécifier un nouveau fichier délimité :
    - Spécifier le Nom à l'étape 1 puis où se trouve le Fichier à l'étape 2.
    - Attention à l'encodage (Windows-1252).
    - Indiquer éventuellement le nombre de lignes à ignorer et si la 1ère ligne comporte des en-têtes.
    - A l'étape 4, spécifier un nom et vérifier les types des colonnes (attention aux dates, ex : "yyyy-MM-dd HH:mm:ss").
    - Rq : Talend indique le type qui correspond le mieux aux données de la colonne, mais il se base sur un échantillon et il se peut qu'il se trompe.

3) Création du job :
- Dans le Repository, créer un nouveau job.
- Choisir les sources et les importer (glisser) dans le Designer avec l'option Input.
- Ajouter le composant tFilterColumns et le relier au fichier customer.csv (clic droit Row/Main) : Configurer le composant tFilterColumns tmappour ne garder que les colonnes utiles.
  Rq : Essayer de donner des noms compréhensibles aux liens et lorsque c'est demandé, récupérer le schéma du composant cible.
- Ajouter le composant tMap et le relier aux autres composants. Dans le tMap :
    - Faites le lien (inner join) entre les deux fichiers customer.csv et state.txt. Rq : les jointures ne se font que du haut vers le bas, l'ordre des sources a donc de l'importance, il n'y a qu'une source main, les autres sont des sources Lookup.
    - Ajouter un filtre pour supprimer les customer dont l'état est inconnu : !cust.CustomerAddress.equals("unknown").
      Rq : on aurait pu utiliser le composant tFilterRow.
    - Ajouter une table cible et relier les colonnes sources aux colonnes cibles et ne garder que les customer de l'état de l'Ohio : state.LabelState.equals("Ohio").
- Ajouter un composant tLogRow (mode Tableau) en output du tMap (clic droit Row).
- Exécuter le job et vérifier le résultat.

- Retourner dans le tMap :
  - Ajouter une seconde cible avec les customer qui n'ont pas passé le 1er filtre : Catch output reject = true, ajouter un tLogRow et vérifier le résultat.
  - Ajouter enfin une troisième cible avec les customer qui n'ont pas passé l'inner join : Catch lookup inner join reject = true, ajouter un tLogRow et vérifier le résultat.



- Ajouter ensuite trois composants correspondant aux formats des cibles (spécifier leurs noms complets) et relier le tMap non plus aux tLogRow mais à ces 3 nouveaux composants. Vérifier les 3 fichiers.
- Afficher un petit commentaire dans le Designer à l'aide du composant Misc / Note.

Aide pour les prochains TP :

- Pour les fichiers Excel, il faut créer une métadonnée par feuille (sauf si les feuilles ont exactement le même schéma).
- Il est possible de glisser un fichier dans le Designer avec l'option Output.
- Connexion à une base de données :
  - Il faut spécifier : Nom, DB Type, Identifiant, Mot de passe, Serveur (localhost), Port (3306), DataBase et vérifier la connexion.
  - Il faut ensuite Récupérer le schéma des tables (clic droit sur la connexion), sélectionner les tables nécessaires.
  - Rq : Pour chaque table, il y a le nom (Db Column) et le type dans la base de données de chaque colonne et la traduction dans Talend (Colonne et Type) ➔ Bien vérifier.
- Dans les composants cibles (pour les BD), vous pouvez spécifier de « vider la table » dans « Action sur la table » pour supprimer les données avant d'en insérer de nouvelles (mais attention aux clés étrangères).
- Dans le tMap, il est possible de créer de nouvelles variables :
  - Exemple âge :

```
Mathematical.INT(TalendDate.formatDate("yyyy",TalendDate.getCurrentDate()))-
Mathematical.INT(TalendDate.formatDate("yyyy",customer.Birth_Date))
```

  - Et de définir des variables complexes :
    - Exemple groupe d'âge :

```
age<30?"<30 years":
age<50?"31-50 years":
">50 years"
```

- On peut de la même manière utiliser diverses fonctions :
  - row.att.equals(" ") ou (row.att == 2 si att de type int)
  - row.att1 + " " + row.att2
  - StringHandling.LEFT(…), RIGHT(…), INDEX(…), LEN()
- Autres composants intéressants :
  - tUniqrow : pour enlever les doublons
  - tUnite : pour faire des unions
  - tDBRow : permet d'exécuter une requête SQL (ex : DELETE)
  - tSplitRow, tDenormalize, tAggregateRow …
- Vous pouvez regrouper les jobs dans un nouveau job (utiliser Trigger / On Component OK).