I. Talend Open Studio

« Talend Open Studio [1] » est un ensemble de produits open source pour le développement, test, déploiement et administration des projets d'intégration de données et d'applications. Talend fournit une plateforme unifiée qui rend la gestion et l'intégration des données et applications plus facile, en fournissant un environnement unifié pour la gestion de tout leur cycle de vie.

Il existe plusieurs solutions offertes par Talend :

- **Big Data** : Environnement qui facilite la gestion des données volumineuses.
- **Data Integration** : Ensemble d'outils pour l'intégration de données pour accéder, transformer et intégrer les données à partir d'un système en temps réel pour remplir les besoins d'intégration des données.
- **Data Quality** : Permet d'assurer le profiling et monitoring des données pour identifier des anomalies et assurer la qualité des données.
- **ESB** : Permet la création, la connexion, la médiation et la gestion des services et leurs interactions.

Pour les besoins de notre TP, nous utilisons « Talend Data Integration » pour la transformation des données et leur intégration. Il est possible de télécharger toutes les solutions de Talend Open Studio sur http://fr.talend.com/products/talend-open-studio

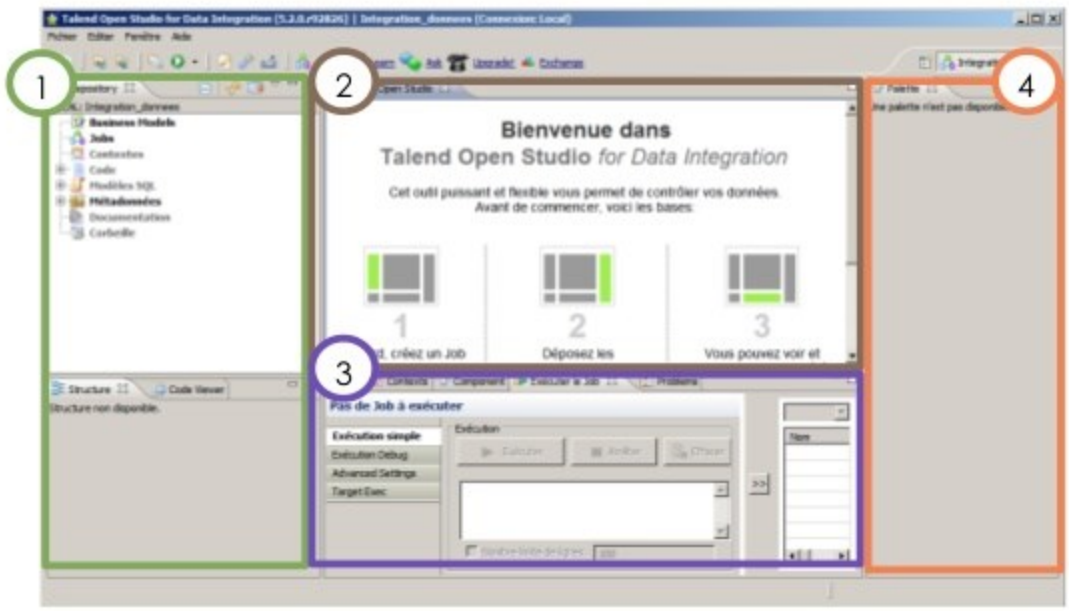
I.1 Installation et Démarrage

- Après avoir installé Talend sur votre machine, le démarrer et créer un nouveau projet intitulé : *Intégration_données*.

Remarque : *Veiller à ce que votre workspace soit à un emplacement accessible en lecture et en écriture (comme vos documents ou votre bureau) : Éviter de le créer directement dans le répertoire d'installation de Talend.*

Après la fermeture de la page de Bienvenue, la fenêtre qui s'affiche aura la forme suivante :

[1] **Talend**: http://fr.talend.com



1	Panneau représentant la structure de votre projet.
2	Panneau affichant l'architecture des Jobs et le code
3	Onglets contenant les propriétés des composants, la console d'exécution, les problèmes…
4	Palette des différents composants disponibles.

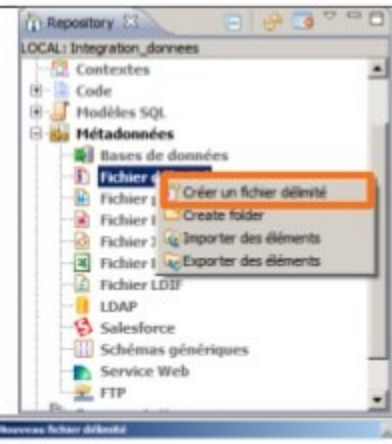
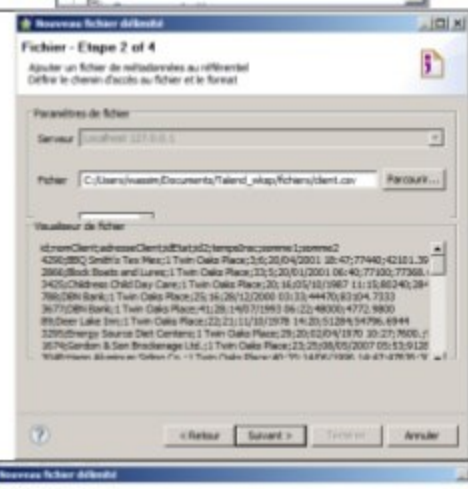
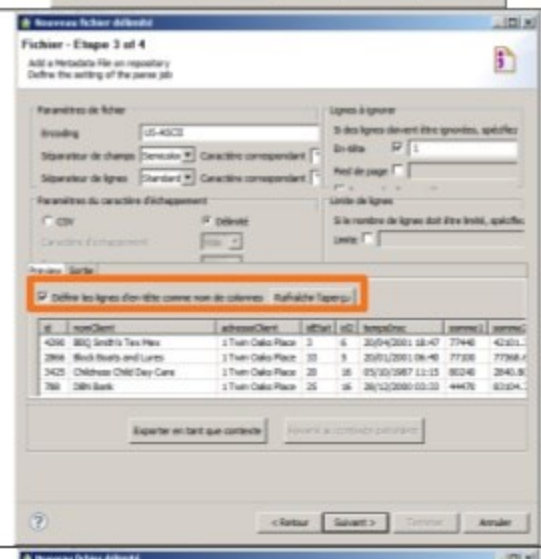
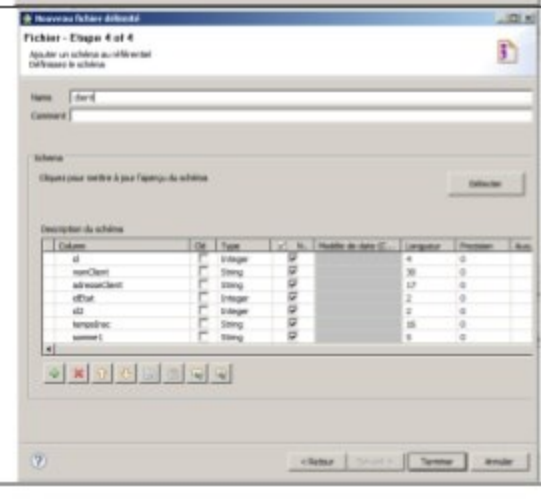
II. Manipulation des Documents

II.1 Préparation des sources de données

Dans ce TP, nous allons manipuler plusieurs sources de données (fichier CSV, fichier texte et base de données) pour en extraire les données, les transformer et les sauvegarder dans d'autres supports. La première étape à réaliser est de définir ces sources de données dans le *Repository* pour pouvoir générer leurs schémas et les utiliser dans les activités suivantes.

Remarque : *Les fichiers que nous allons ajouter (client.csv et etats.txt) vous ont été fournis avec le support de TP.*

Pour faire cela, suivre les étapes suivantes :

Dans le panneau (1) représentant le *Repository*, développer la section *Métadonnées*. Pour définir des sources sous forme de champs séparés par des délimiteurs (comme des fichiers csv ou texte), choisir : *Créer un fichier délimité*. Entrer le nom du fichier dans la fenêtre qui apparaît : *client* (dans notre cas, nous allons ajouter le fichier *client.csv*)	
Choisir ensuite le fichier que vous désirez ajouter. Naviguer pour cela vers le fichier *client.csv* qui vous a été fournis. Le visualiseur de fichier vous permet d'avoir une idée sur le contenu de ce fichier. Cliquer sur *suivant*.	
Dans la fenêtre suivante, cliquer sur la case *Définir les lignes d'en-tête comme nom de colonne*. Cliquer ensuite sur *Rafraîchir l'aperçu*. L'aperçu du fichier extrait sera mis à jour, de manière à ce que la première ligne du fichier représente les noms des champs. Cliquer sur *suivant*.	
Modifier le nom du schéma du fichier délimité (*client*), et observer la composition des champs dans le panneau *Description du schéma*. Vous pourrez ainsi modifier les données du schéma à votre guise. Dans notre cas, **ne pas oublier de cocher la case *Clé* pour le champ *id*.** Vous pourrez également modifier les longueurs des champs (les valeurs par défaut ont été calculées par Talend selon les données déjà présentes dans le fichier). Cliquer sur *terminer*.	

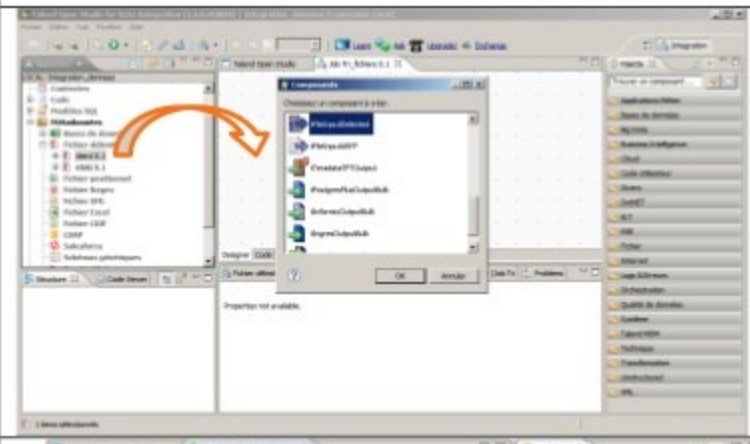
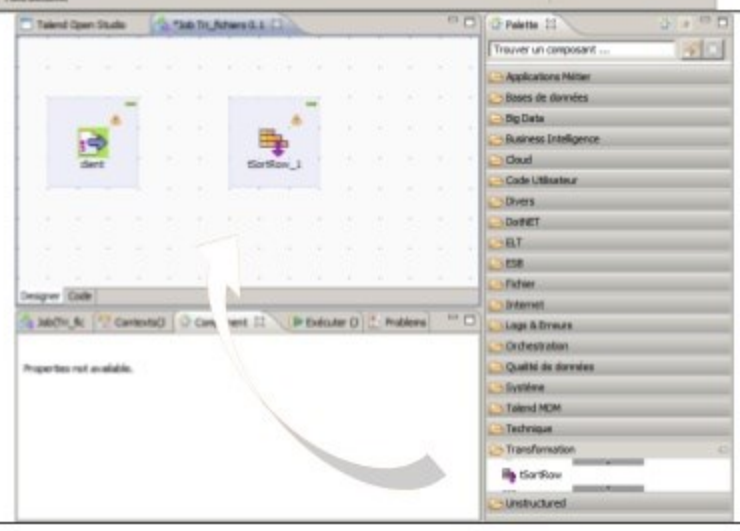
Vous avez ainsi ajouté un fichier source, dont le schéma pourra être utilisé dans toute l'application.

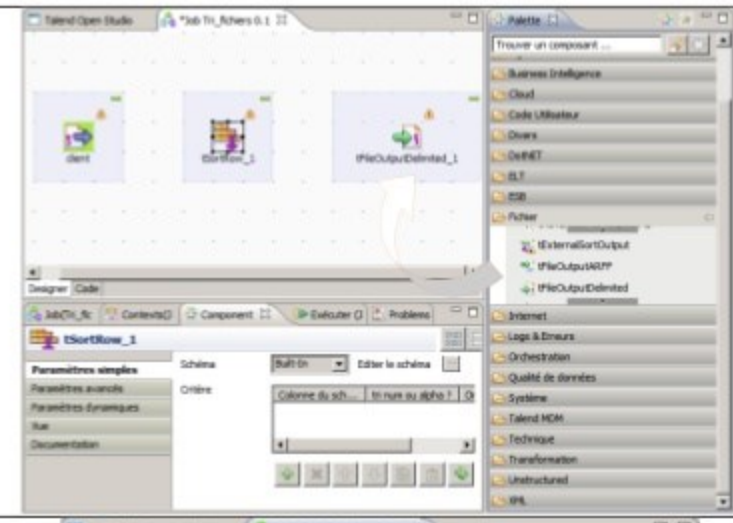
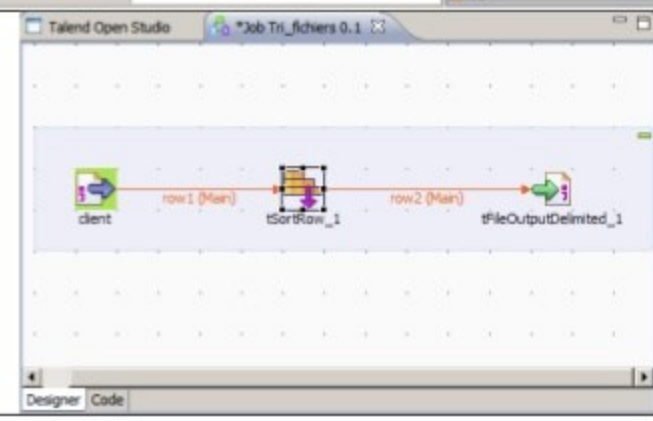
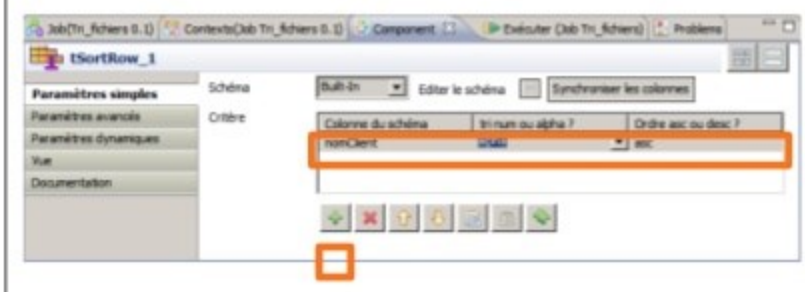
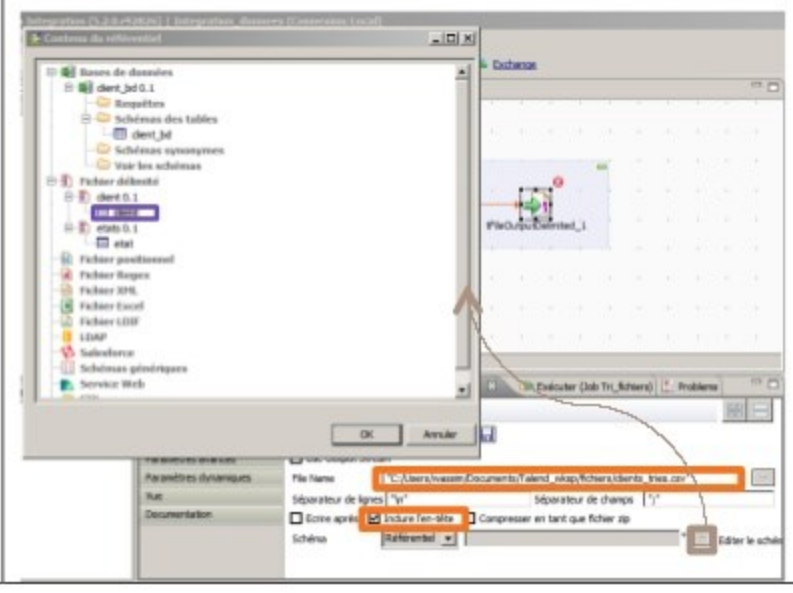
Activité 1.

- Générer de même le schéma du fichier *état.txt* qui vous est fourni.
- Dans le SGBD de votre choix, créer une base de données *client_bd*, contenant une table appelée *client*. La structure de cette table n'a pas d'importance, elle sera écrasée plus tard.
- Ajouter la base de données comme source dans la partie *Métadonnées*, et ajouter la table *client* aux schémas des tables.

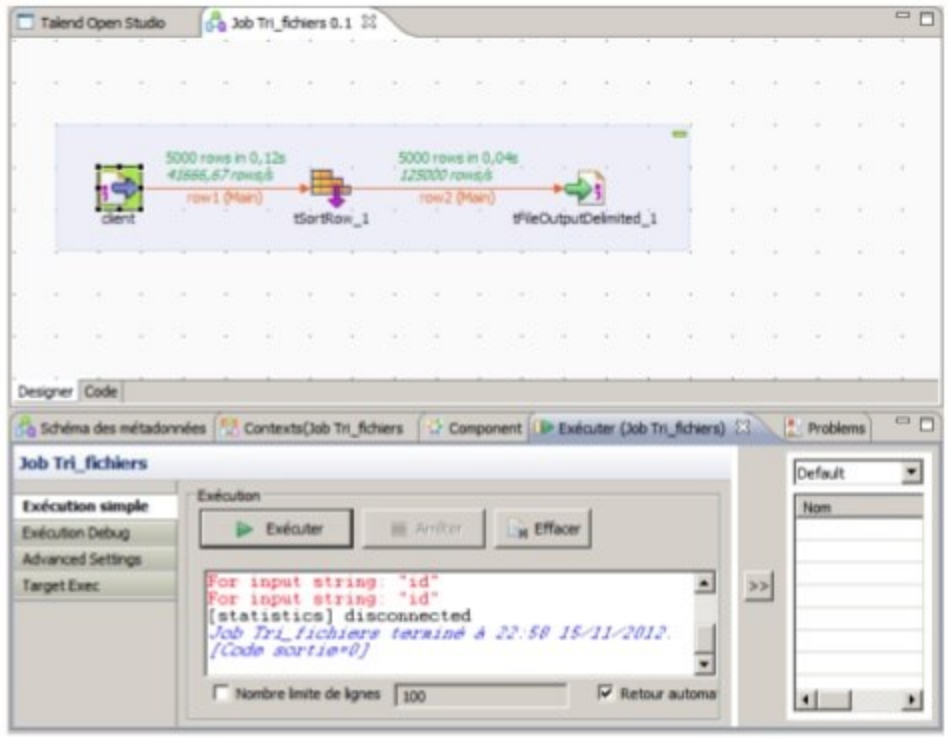
II.2 Tri de documents

Dans cette première activité, on se propose de trier le contenu du fichier *client.csv* de manière automatique, en utilisant les composants Talend. Pour cela, suivre les étapes suivantes :

Créer un nouveau Job que vous appellerez Tri_fichiers	
Glisser le fichier délimité *client 0.1*, que vous avez créé précédemment, dans le panneau (2). Indiquer dans la fenêtre qui apparaît que c'est un *tFileInputDelimited*. Cliquer sur *OK*.	
Dans le panneau (4), représentant la palette, choisir le composant *tSortRow* dans la catégorie *Transformation*. Ce composant permet, comme son nom l'indique, de faire le tri d'un ensemble de données, selon une colonne particulière. Faire glisser ce composant dans la fenêtre principale.	

Pour représenter le fichier de sortie, faire glisser le composant *tFileOutputDelimited* dans la fenêtre principale. Il se trouve sous la catégorie *Fichier -> Ecriture*.	
Relier les trois éléments pour représenter la chaîne d'exécution. Pour cela, faire un clic droit sur le composant *client*, maintenir enfoncé, et glisser vers le composant de tri. Faire de même entre le composant de tri et le fichier de sortie.	
Nous allons maintenant configurer les trois composants. Nous allons d'abord définir le nom du client comme critère de tri, par ordre alphabétique, du fichier source. Cliquer sur le composant de tri. Sous l'onglet *Composant* du panneau (3), cliquer sur (+). Modifier la valeur des champs insérés, pour faire le tri selon le nom de client, par ordre alphabétique ascendant.	
Cliquer ensuite sur le composant de sortie. - Choisir l'emplacement où on désire sauvegarder le fichier de sortie - Cocher la case : *Inclure l'en-tête* pour que l'en-tête des colonnes s'affiche dans le fichier de sortie - Devant la case *Schéma*, changer le type de schéma vers *Référentiel*, puis cliquer sur [...] à côté de *Editer le schéma*. Cela permettra de définir la structure des champs du fichier de sortie. - Dans la fenêtre affichée, choisir le schéma *client* du fichier délimité que vous avez créé.	

Une fois ces étapes terminées, enregistrer le projet. Pour exécuter le processus, Cliquer sur l'onglet *Exécuter* du panneau (3), puis cliquer sur *Exécuter*. Ou alors taper *F6*. A la fin de l'exécution, la trace suivante est affichée sur la fenêtre principale:



Vérifier que le fichier trié a bien été créé dans le répertoire que vous avez spécifié plus tôt.

Activité 2.

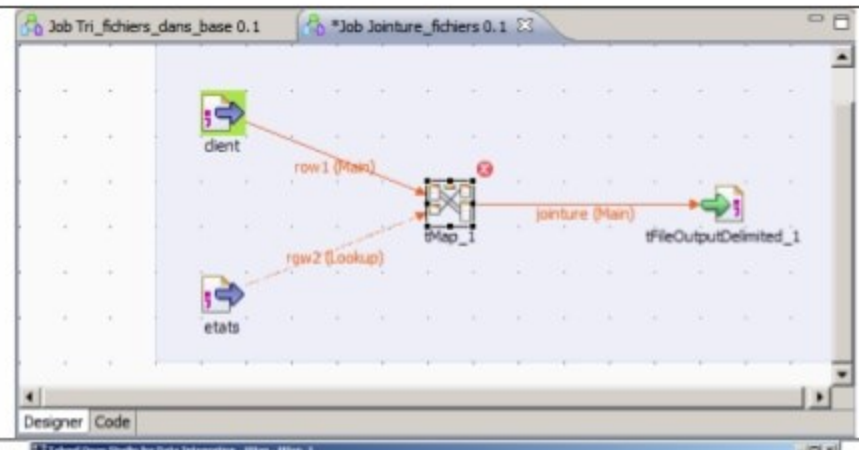
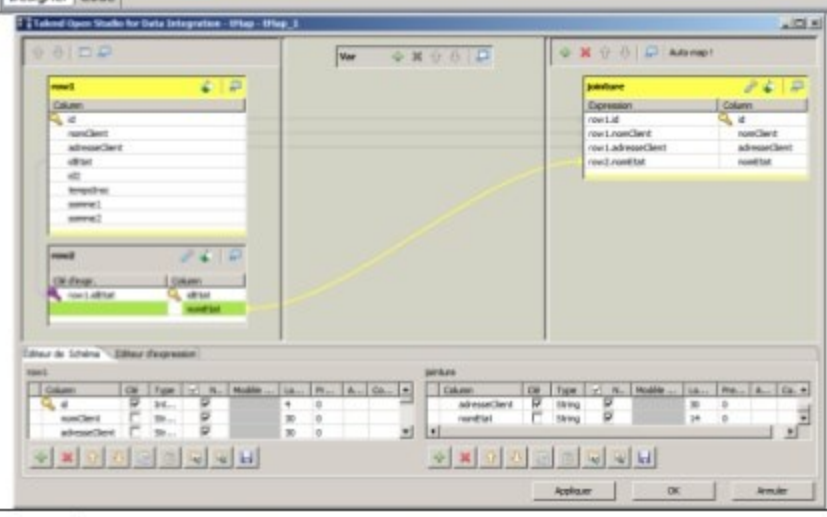
- Dupliquer le job *Tri_fichiers* et le nommer *Tri_fichier_dans_base*
- Copier les données générées dans le fichier délimité de sortie dans la base de données *client_bd* que vous avez créé dans l'activité précédente (au lieu d'un fichier CSV). A la création, la table cible sera écrasée et remplacée par la table contenant les données triées.

II.3 Jointure de fichiers

Le fichier *etat.txt* permet d'associer l'identifiant d'un état américain avec le nom de cet état. On se propose de faire la jointure des fichiers *client.csv* et *etat.txt* pour remplacer l'identifiant de l'état dans les données du client par son nom.

Pour faire cela, créer un nouveau Job *Jointure_fichiers* et suivre les étapes suivantes :

Glisser les deux fichiers délimités *client* et *etat* dans le panneau principal.
Glisser le composant *tMap*, de la catégorie *Transformation* dans le panneau principal. Ce composant permet de transformer et diriger les données à partir d'une ou plusieurs sources vers une ou plusieurs destinations.
Enfin, faire glisser un fichier délimité

de sortie.	
Relier les différents composants. Relier le fichier d'entrée *client* d'abord à la *tMap*, puis le fichier *etats*. Relier enfin le la *tMap* vers le fichier de sortie. Appeler la sortie *jointure*.	
Double cliquer sur la *tMap* pour la configurer. Une fenêtre s'ouvre. Commencer par relier le champ *idEtat* de la première table *row1*, au champ *idEtat* de la table *row2*. Faire glisser ensuite les champs *id, nomClient,* et *adresseClient* de *row1*, puis *nomEtat* de *row2* vers la table de destination *jointure*.	

Configurer ensuite le fichier de sortie en précisant son chemin , et en incluant l'en-tête.
Exécuter le Job, et vérifier le fichier de sortie.

Activité 3.

- Créer un nouveau Job *Jointure_Tri_fichiers_de_base*
- Ce job permet de :
  - faire la jointure entre la table *client* créée dans l'activité 2 et le fichier *etat.txt* pour obtenir les champs *id, nomClient, adresseClient* et *nomEtat*.
  - trier ces données jointes par **nom d'état**, avant de les stocker dans un fichier texte *clients-etat.txt* dont les champs sont délimités par le caractère « | ».

II.4 Sélection des données

Il est possible de filtrer les données, en rejetant par exemple les entrées erronées. On peut remarquer dans les données du fichier *client.csv* que certaines entrées ne comportent pas de nom d'état. On désire filtrer ces données, et n'enregistrer dans le fichier de sortie que les données comportant un nom d'état. Les autres données pourront être affichées dans la console.