



Compte rendu des travaux pratiques : TP2 et TP3

Elaboré par

Houssem Eddine Hamzaoui

Section

IGL3

Encadré par

Mme Manel Zekri

Année Universitaire

2024 - 2025

TP2 : Créer une métadonnée de connexion à un fichier délimité

Introduction

Dans les processus d'intégration de données, la gestion des schémas de fichiers délimités est essentielle pour structurer les informations provenant de sources diverses. Ce TP se concentre sur l'utilisation de Talend Open Studio pour configurer des métadonnées adaptées à des fichiers délimités comme customer.csv et state.txt. Ces métadonnées permettent une manipulation standardisée et réutilisable des données dans des jobs Talend.

Objectif

Ce TP vise à :

- 1. Créer des connexions de métadonnées pour deux fichiers délimités (customer.csv et state.txt) dans le référentiel Talend.**
- 2. Paramétrer les schémas pour structurer les données.**
- 3. Acquérir une méthode de travail facilitant la réutilisation des connexions dans plusieurs jobs.**

Documentation des étapes

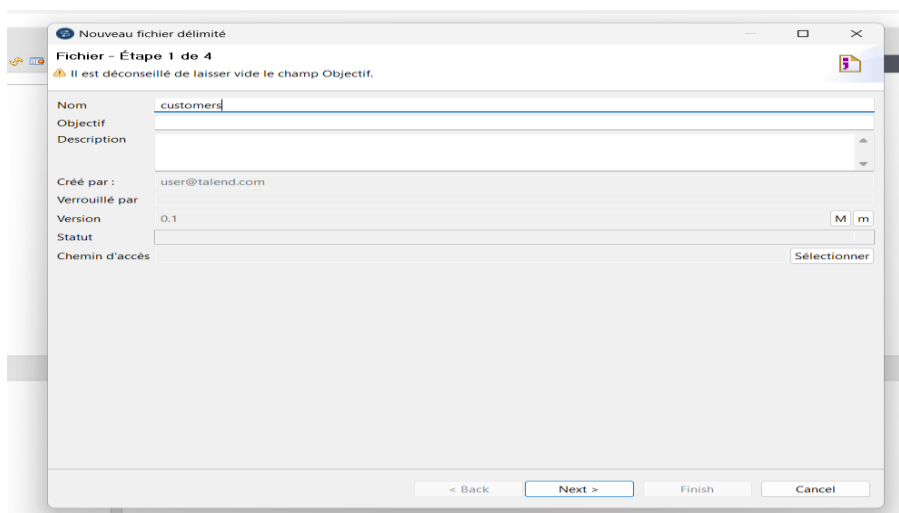
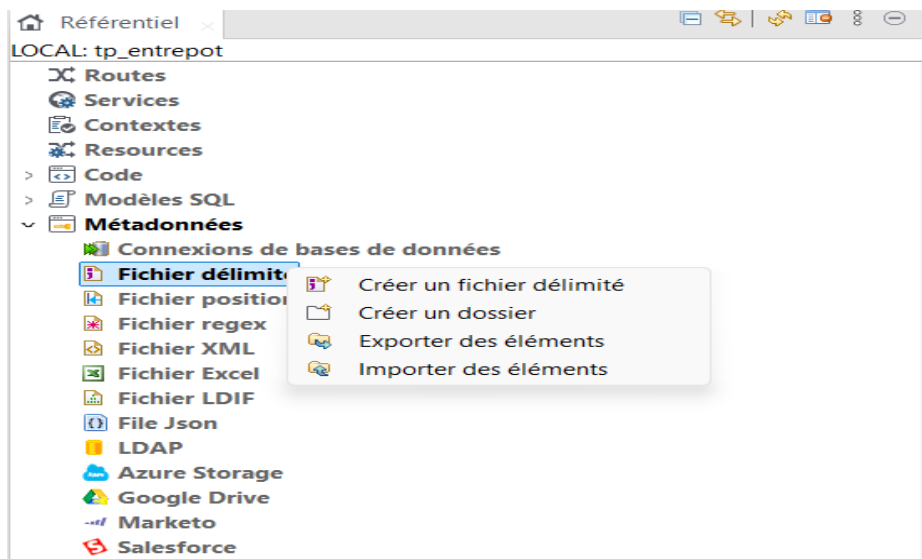
Étape 1 : Créer la métadonnée *customers*:

1) Créer un nœud dans Metadata :

Dans Talend, développez le nœud Metadata > File Delimited.

Faites un clic droit et sélectionnez Create file delimited. Cela lance un assistant.

Donnez un nom significatif à la métadonnée : *customers*



2) Importer le fichier source :

- Utilisez le bouton **Browse** pour sélectionner le fichier **customer.csv**.
- Spécifiez le format selon votre système d'exploitation (Windows, Linux, etc.).

Nouveau fichier délimité

Fichier - Étape 2 de 4

Ajouter un fichier de métadonnées au référentiel
Définissez le chemin d'accès au fichier et les paramètres de format

Paramètres du fichier

Serveur: Localhost 127.0.0.1

Fichier: C:/Users/hamza/OneDrive/Bureau/tp_entrepot/tp2/customer (2).csv

Format: WINDOWS

Visualiseur de fichier

```
*****/*****  
/***** Extract on Mon Oct 02 10:30:19 CEST 2006 *****/*****  
*****/*****  
  
id;CustomerName;CustomerAddress;idState;id2;RegTime;RegisterTime;Sum1;Sum2  
1;Griffith Paving and Sealcoat;talend@apres91;7;41;03/11/2006 09:20;2001-01-17 06:26:40.000;67852;61521.4852  
2;Bill's Dive Shop;511 Maple Ave. Apt. 1B;35;5;19/11/2004 15:48;2002-06-07 09:40:00.000;88792;15434.1000  
3;Childress Child Day Care;662 Lyons Circle;1;28;16/02/2005 08:27;1990-04-01 21:00:00.000;35340;17856.8818  
4;Facelift Kitchen and Bath;unknown;0;15;22/08/2002 09:55;1972-04-23 18:00:00.000;6097;55560.2387  
5;Terrinni & Son Auto and Truck;770 Exmoor Rd.;0;9;28/06/2001 09:15;1982-04-19 10:26:40.000;5146;39098.1148  
6;Kermit the Pet Shop;1860 Parkside Ln.;28;15;17/08/2003 10:07;2006-05-27 17:00:00.000;16087;29924.9294
```

< Back Next > Finish Cancel

3) Paramétrer le schéma :

- **En-têtes** : Cochez **Set heading row as column names** pour récupérer les noms des colonnes automatiquement.
- **Lignes ignorées** : Spécifiez que les 5 premières lignes (Header) doivent être ignorées.
- **Structure des colonnes** : Configurez chaque colonne en précisant son nom, son type (String, Integer, etc.), et sa longueur.

Nouveau fichier délimité

Fichier - Étape 3 de 4

Ajouter un fichier de métadonnées au référentiel
Définissez les paramètres du Job de passage

Paramètres du fichier

Encodage: US-ASCII

Séparateur de champs: Semi

Séparateur de lignes: Stan

Paramètres du caractère d'échappement

CSV

Caractère d'échappement: Vider

Entourage du texte: Vider

Scinder la ligne avant le champ

Lignes à ignorer

Si des lignes doivent être ignorées, spécifiez les paramètres

En-tête: ☒ 6

Pied de page: ☐

Ignorer les lignes vides: ☐

Limite de lignes

Si le nombre de lignes doit être limité, spécifier ce nombre

Limite: ☐

Aperçu

☒ Définir la ligne d'en-tête comme nom de colonnes

Actualiser l'aperçu

id	CustomerName	CustomerAddress	idState	id2	RegTime	RegisterTime	Sum
1	Griffith Paving and Sealcoat	talend@apres91	7	41	03/11/2006 09:20	2001-01-17 06:26:40.000	67852
2	Bill's Dive Shop	511 Maple Ave. Apt. 1B	35	5	19/11/2004 15:48	2002-06-07 09:40:00.000	88792
3	Childress Child Day Care	662 Lyons Circle	1	28	16/02/2005 08:27	1990-04-01 21:00:00.000	35340

Exporter en tant que contexte Revenir au contexte précédent

< Back Next > Finish Cancel

4) Validation :

- **Rafraîchissez l'aperçu des données pour vérifier la cohérence avec le fichier source.**
- **Cliquez sur Finish pour enregistrer.**

Nouveau fichier délimité

Fichier - Étape 4 de 4

Ajouter un schéma au référentiel
Définissez le schéma

Nom

Commentaire

Schéma

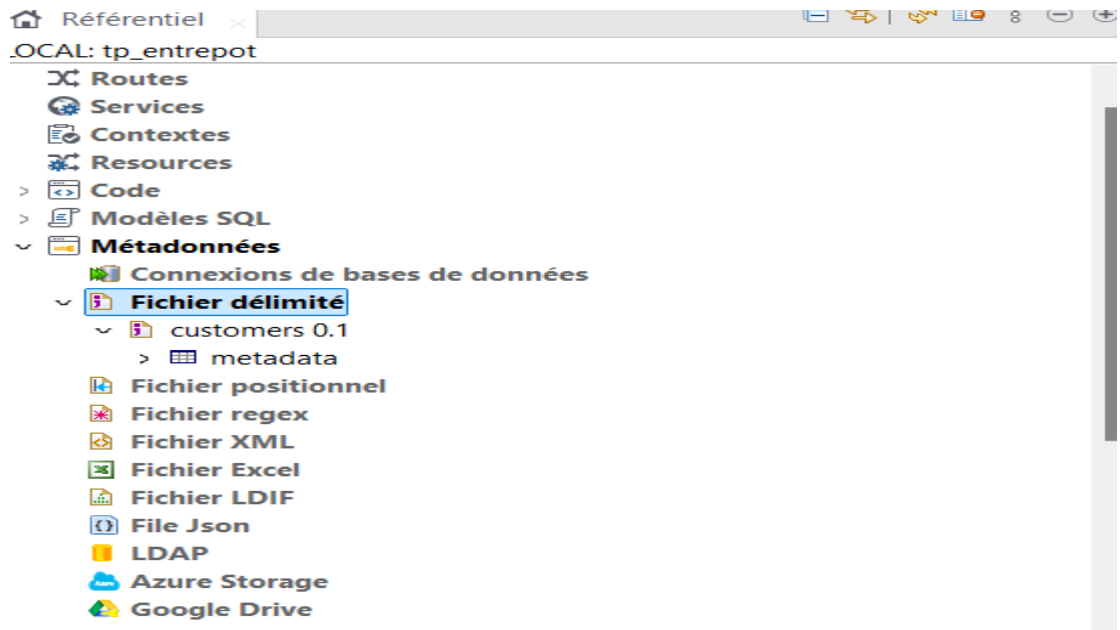
Cliquez pour mettre à jour la prévisualisation du schéma.

Description du schéma

Colonne	Clé	Type	<input checked="" type="checkbox"/> N.	Modèle de date ...	Longueur	Précision	Par dé...	Commen...
id	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		2	0		
CustomerName	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		30	0		
CustomerAddress	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		25	0		
idState	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		2	0		
id2	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		2	0		
RegTime	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		16	0		
RegisterTime	<input type="checkbox"/>	Date	<input checked="" type="checkbox"/>	"dd-MM-yyyy"	23	0		
Sum1	<input type="checkbox"/>	Float	<input checked="" type="checkbox"/>			0		

< Back Next > Finish Cancel

On trouve les métadonnées sous fichier délimité :



Étape 2 : Créer la métadonnée *states*

1. Répétez les étapes précédentes :

- Suivez le même processus pour créer une métadonnée pour *state.txt*.
- Donnez le nom *states* à cette connexion.
- Importez le fichier, configurez les en-têtes et les colonnes, et validez le schéma.

1)

Nouveau fichier délimité

Fichier - Étape 1 de 4

⚠ Il est déconseillé de laisser vide le champ Objectif.

Nom: states

Objectif:

Description:

Créé par: user@talend.com

Verrouillé par:

Version: 0.1

Statut:

Chemin d'accès: Sélectionner

< Back Next > Finish Cancel

2)

Nouveau fichier délimité

Fichier - Étape 2 de 4

Ajouter un fichier de métadonnées au référentiel
Définissez le chemin d'accès au fichier et les paramètres de format

Paramètres du fichier

Serveur: Localhost 127.0.0.1

Fichier: C:/Users/hamza/OneDrive/Bureau/tp_entrepot/tp2/state (1).txt

Format: WINDOWS

Parcourir...

Visualiseur de fichier

idState;LabelState
1;Alabama
2;Alaska
3;Arizona
4;Arkansas
5;California
6;Colorado
7;Connecticut
8;Delaware
9;Florida
10;Georgia
11;Hawaii

< Back Next > Finish Cancel

3)

Nouveau fichier délimité

Fichier - Étape 3 de 4

Ajouter un fichier de métadonnées au référentiel
Définissez les paramètres du Job de parsing

Paramètres du fichier

Encodage: US-ASCII

Séparateur de champs: Semi Caractère correspondant: ";"

Séparateur de lignes: Stan Caractère correspondant: "\n"

Paramètres du caractère d'échappement

☐ CSV ☒ Délimité

Caractère d'échappement: Vider

Entourage du texte: Vider

☐ Scinder la ligne avant le champ

Lignes à ignorer

Si des lignes doivent être ignorées, spécifiez les paramètres s

En-tête: ☒ 1

Pied de page: ☐

☐ Ignorer les lignes vides

Limite de lignes

Si le nombre de lignes doit être limité, spécifiez ce nombre

Limite: ☐

Aperçu | Sortie

☒ Définir la ligne d'en-tête comme nom de colonnes Actualiser l'aperçu

idState	LabelState
1	Alabama
2	Alaska
3	Arizona
4	Arkansas

Exporter en tant que contexte Revenir au contexte précédent

< Back Next > Finish Cancel

4)

Nouveau fichier délimité

Fichier - Étape 4 de 4

Ajouter un schéma au référentiel
Définissez le schéma

Nom: metadata

Commentaire:

Schéma

Cliquez pour mettre à jour la prévisualisation du schéma. Détecter

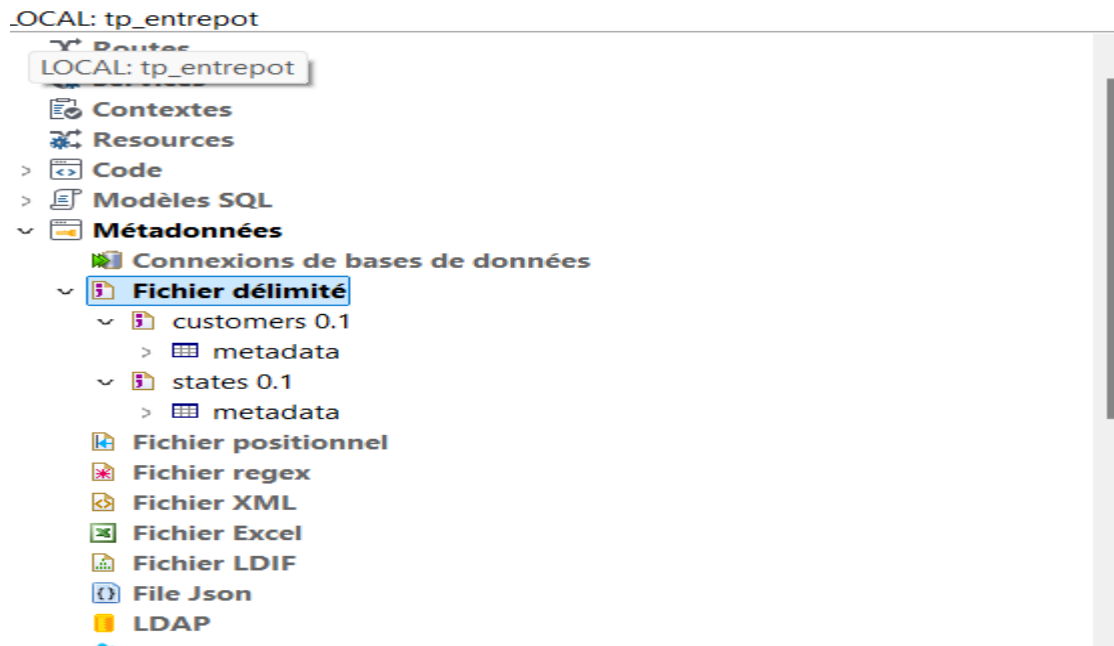
Description du schéma

Colonne	Clé	Type	<input checked="" type="checkbox"/> N.	Modèle de date (...)	Longueur	Précision	Par déf...	Commen...
idState	<input checked="" type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		2	0		
LabelState	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		14	0		

+ ✖ ⬆ ⬇ 📄 🗑 🔄 🔄

< Back Next > Finish Cancel

On trouve les métadonnées sous fichier délimité :



Résultat final :

- Les métadonnées customers et states apparaissent sous le nœud Metadata > File Delimited. Elles sont prêtes à être réutilisées dans des jobs Talend.

Conclusion

Ce TP met en évidence l'importance d'une gestion centralisée des métadonnées pour garantir une structure homogène des données. La réutilisation des schémas facilite les tâches d'intégration ultérieures, réduisant les risques d'erreur.

TP3 : Créer une jointure entre deux fichiers

Introduction

Les opérations de jointure sont des étapes fondamentales dans la gestion de données. Ce TP utilise le composant tMap pour combiner les données de deux fichiers (customer.csv et state.txt) en un fichier final customers+states.csv. Ce fichier fusionné regroupe les informations client et leur état, illustrant la puissance de transformation des données dans Talend.

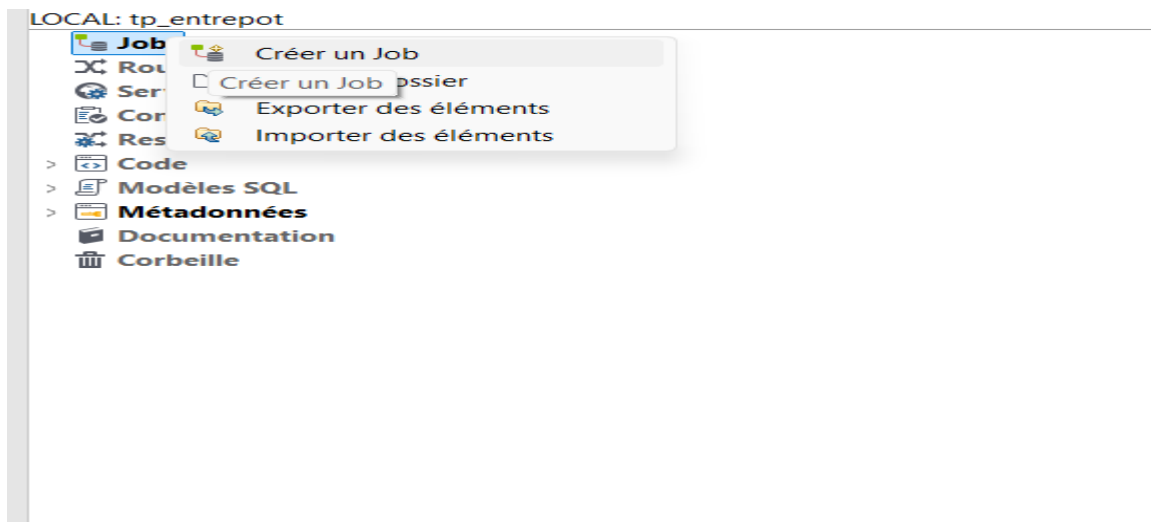
Objectif

L'objectif principal est de construire un Job capable de :

1. **Charger deux fichiers délimités en entrée.**
2. **Réaliser une jointure logique sur la colonne idState.**
3. **Produire un fichier combiné contenant des données consolidées pour des analyses.**

Documentation des étapes

Étape 0: créer le job:



Étape 1 : Configuration des fichiers d'entrée

Ajout des composants d'entrée :

- Dans le Job Designer, ajoutez deux composants tFileInputDelimited.



- Paramétrez chaque composant pour utiliser les métadonnées créées dans le TP2 (customers et states).

Dans la vue Component :

Pour spécifier les propriétés du composant, sélectionnez Repository (Référentiel) dans la liste Property ,Type (type de propriété) puis cliquez sur le bouton [...] situé à côté du champ Edit schema pour vérifier le schéma du fichier. L'assistant Edit parameter using repository s'ouvre.

Type de propriété **Référentiel** DELIM:customers ...

Schéma **Référentiel** DELIM:customers - metadata * ... Modifier le schéma ...

"When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."

Nom de fichier/Flux "C:/Users/hamza/OneDrive/Bureau/tp_entrepot/tp2/customer (2).csv" *

Séparateur de lignes "\n" *

Séparateur de champs "," *

☐ Options CSV

En-tête 6

Pied de page 0

Tfileinputdelimited 1

Job(HowToSetupJoinLink 0.1) Contexts(HowToSetupJoinLink) Composant Exécuter (Job HowToSetupJoinLink)

tFileInputDelimited_2

Paramètres simples Type de propriété **Référentiel** DELIM:states ...

Paramètres avancés Schéma **Référentiel** DELIM:states - metadata * ... Modifier le schéma ...

Paramètres dynamiques "When the input source is a stream or a zip file, footer and random shouldn't be bigger than 0."

/ue Nom de fichier/Flux "C:/Users/hamza/OneDrive/Bureau/tp_entrepot/tp2/state (1).txt" *

Documentation Séparateur de lignes "\n" *

Séparateur de champs "," *

☐ Options CSV

En-tête 1

Pied de page 0

Tfileinputdelimited 2

Vérifiez les schémas pour garantir leur cohérence avec les fichiers sources.

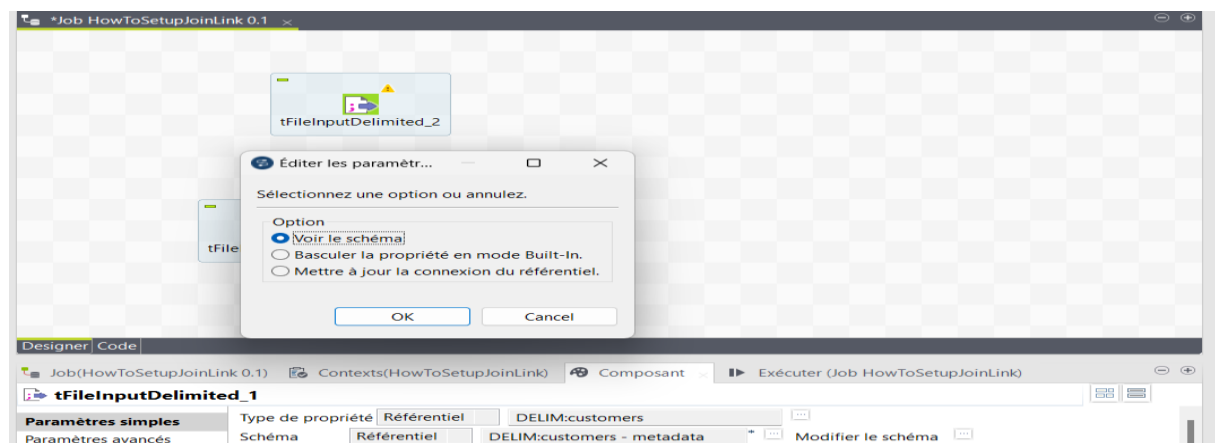


Schéma de tFileInputDelimited_1

Colonne utilisée	Colonne	Clé	Type	N.	Modèle de date (...)	Longueur	Précision	Par déf...	Comment...
<input checked="" type="checkbox"/>	id	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		9	0		
<input checked="" type="checkbox"/>	CustomerName	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		255	0		
<input checked="" type="checkbox"/>	CustomerAddress	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		255	0		
<input checked="" type="checkbox"/>	idState	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		2	0		
<input checked="" type="checkbox"/>	id2	<input type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		2	0		
<input checked="" type="checkbox"/>	RegTime	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		30	0		
<input checked="" type="checkbox"/>	RegisterTime	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		30	0		
<input checked="" type="checkbox"/>	Sum1	<input type="checkbox"/>	Float	<input checked="" type="checkbox"/>		10	0		
<input checked="" type="checkbox"/>	Sum2	<input type="checkbox"/>	Float	<input checked="" type="checkbox"/>		10	5		

OK Cancel

Shéma de tfileinputdelimited1

Schéma de tFileInputDelimited_2

Colonne utilisée	Colonne	Clé	Type	N.	Modèle de date (C...	Longueur	Précision	Par déf...	Comment...
<input checked="" type="checkbox"/>	idState	<input checked="" type="checkbox"/>	Integer	<input checked="" type="checkbox"/>		2	0		
<input checked="" type="checkbox"/>	LabelState	<input type="checkbox"/>	String	<input checked="" type="checkbox"/>		14	0		

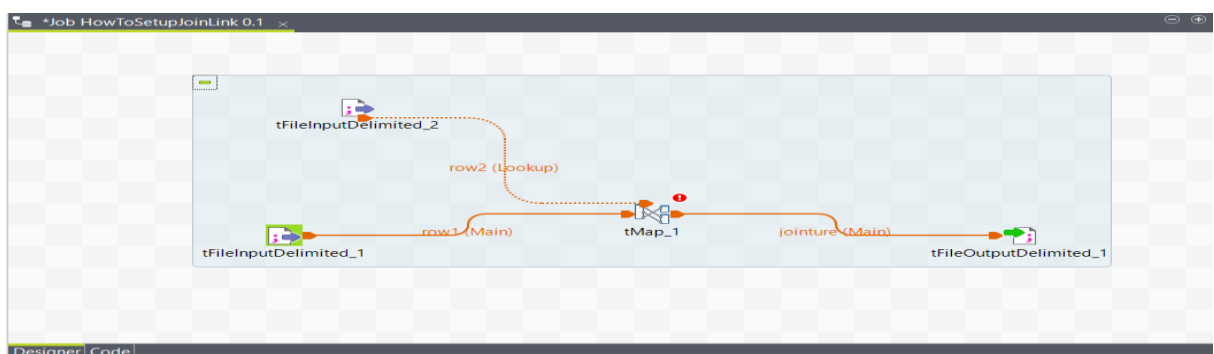
OK Cancel

Shéma de tfileinputdelimited2

Étape 2 : Ajouter le composant de transformation

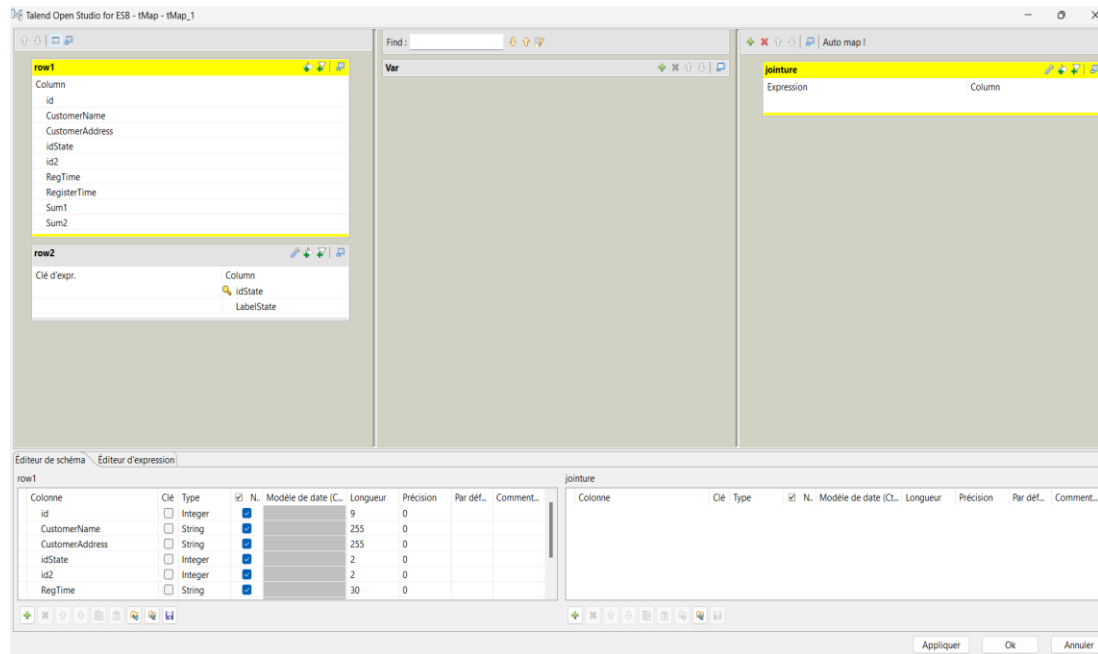
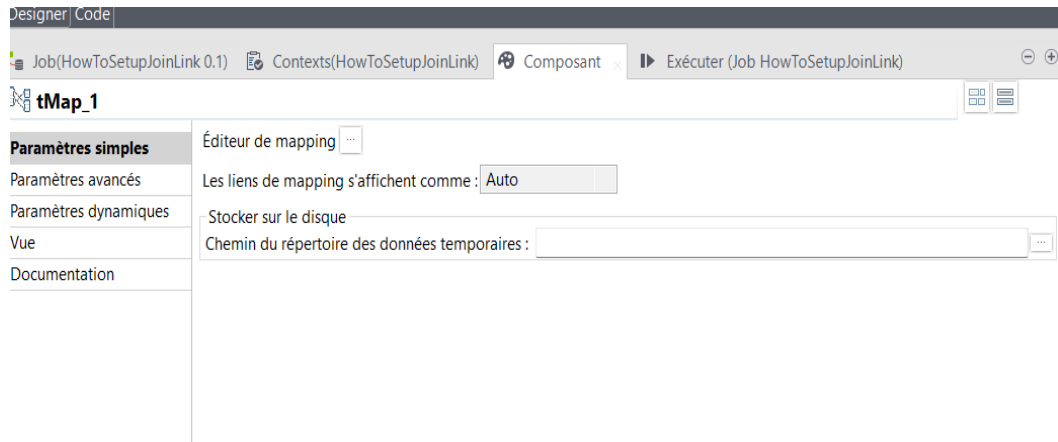
1. Insertion du composant tMap :

- Ajoutez le composant tMap dans le Job Designer.
- Reliez chaque fichier d'entrée au tMap via des liens Row > Main.

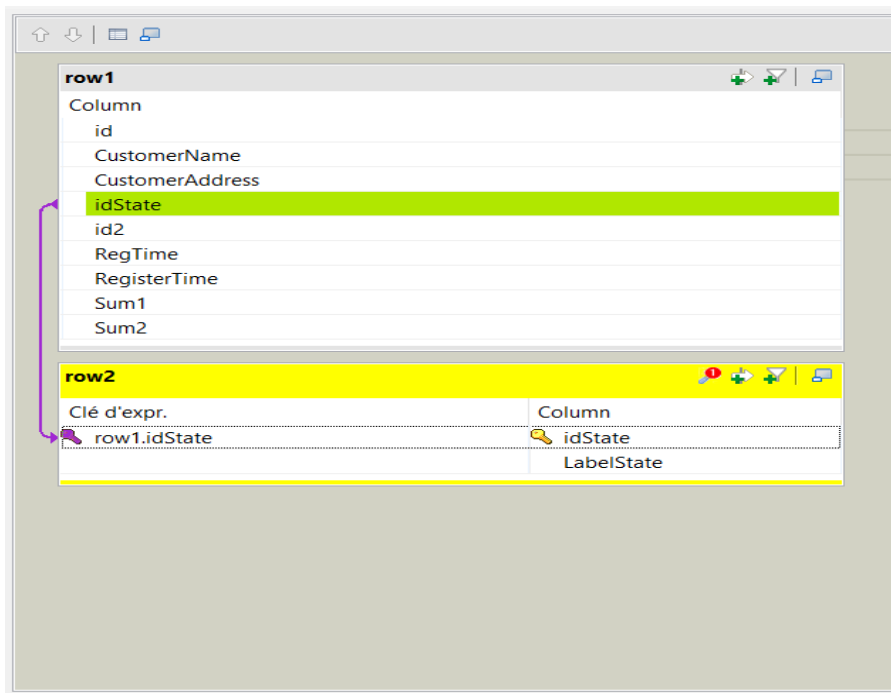


2. Configurer la jointure dans le tMap :

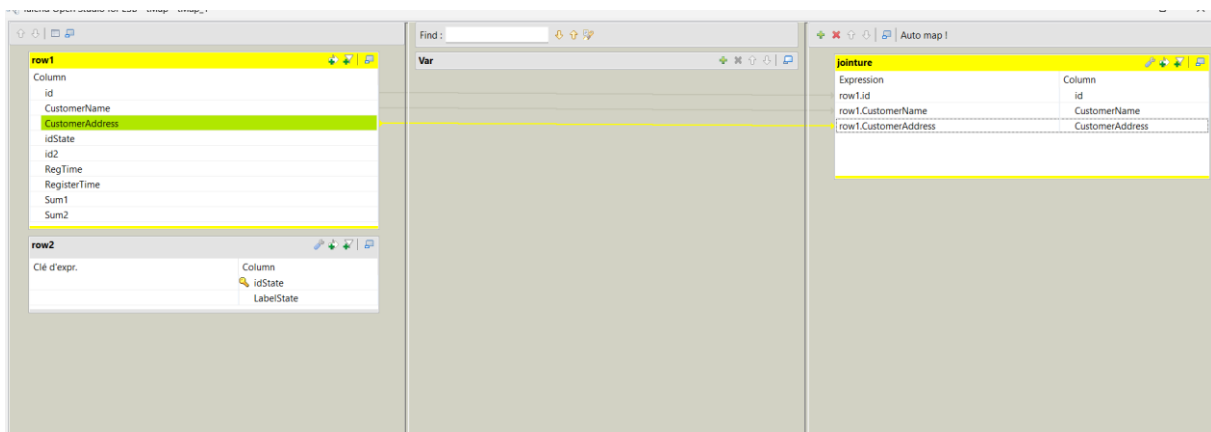
- Ouvrez l'éditeur du tMap.



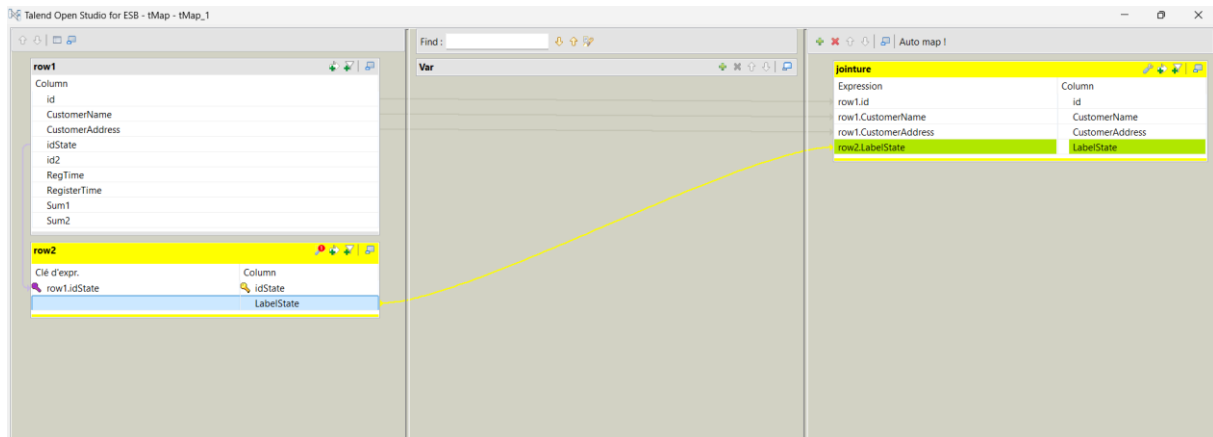
- **Colonne de jointure : Faites correspondre idState entre les deux fichiers.**



- **Colonnes de sortie :**
 - **Depuis customer.csv : Ajoutez id, CustomerName, et CustomerAddress.**

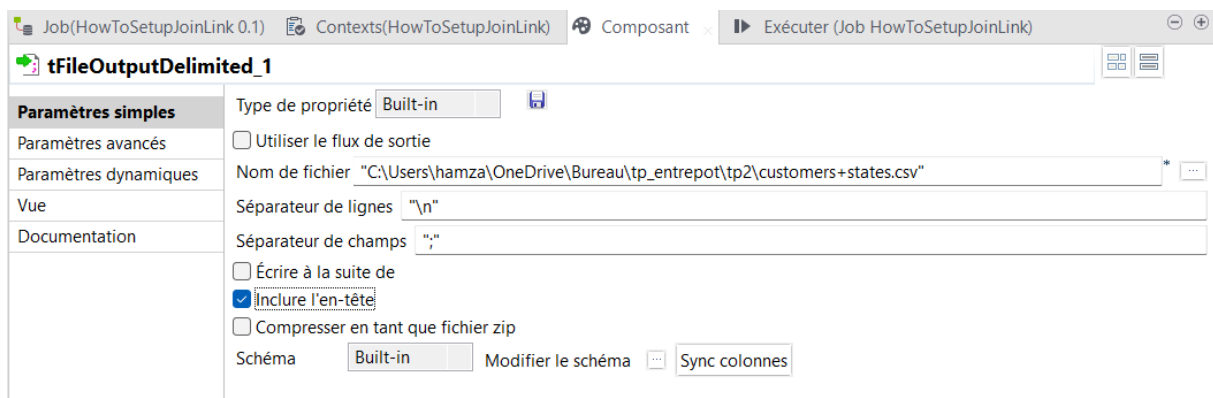


- **Depuis state.txt : Ajoutez LabelState.**



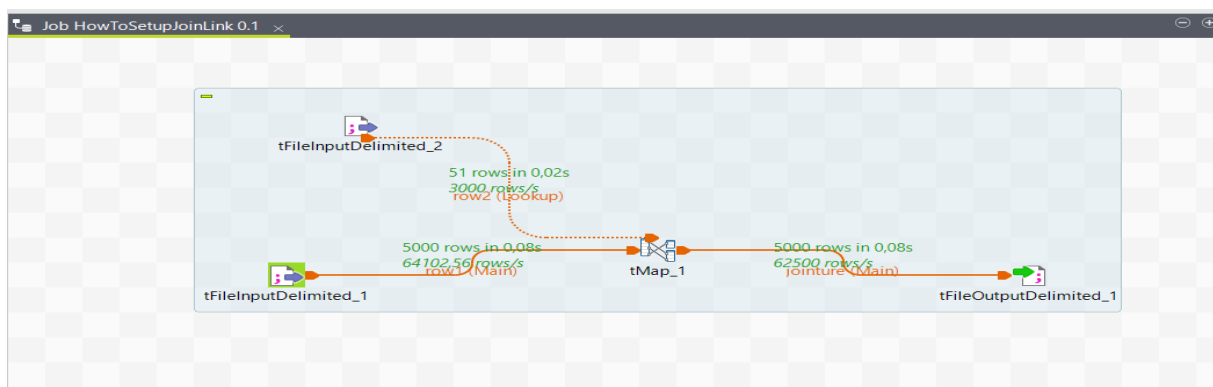
Étape 3 : Configuration du fichier de sortie

1. **Ajout du composant de sortie :**
 - Ajoutez un composant **tFileOutputDelimited**.
 - Spécifiez le nom et l'emplacement du fichier de sortie : customers+states.csv.
2. **Options supplémentaires :**
 - Activez **Include Header** pour inclure les noms des colonnes dans le fichier final.

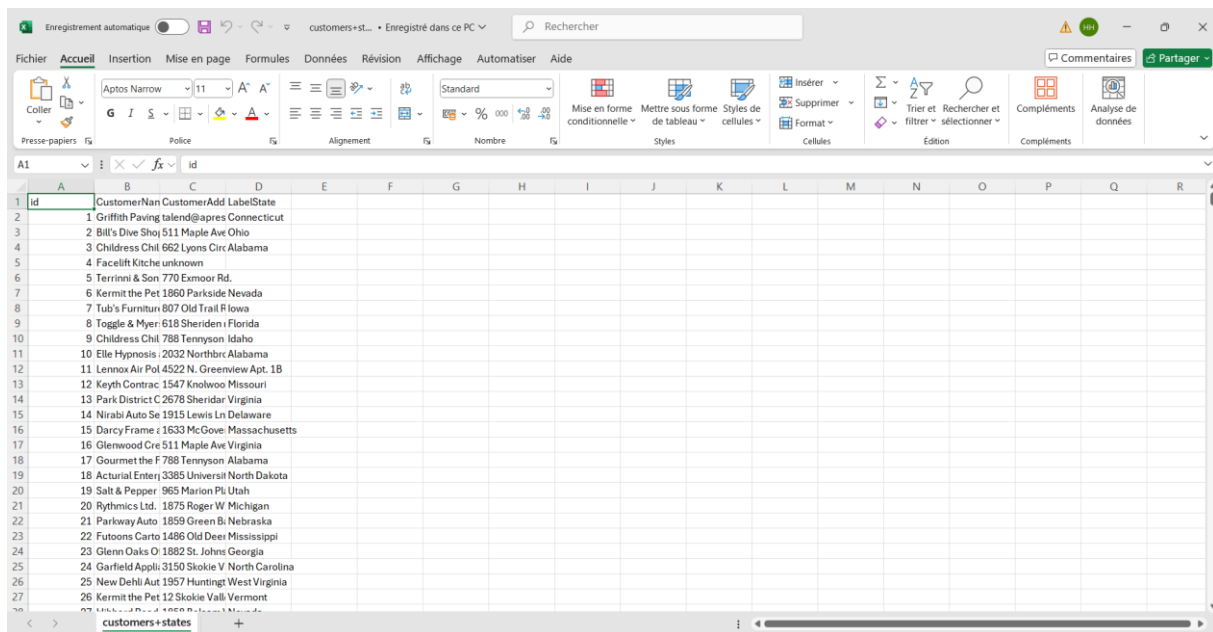


Étape 4 : Exécution du Job

1. **Sauvegarde et exécution :**
 - Enregistrez le job et lancez l'exécution (Ctrl+S, puis F6).
 - Vérifiez la création et le contenu du fichier customers+states.csv.

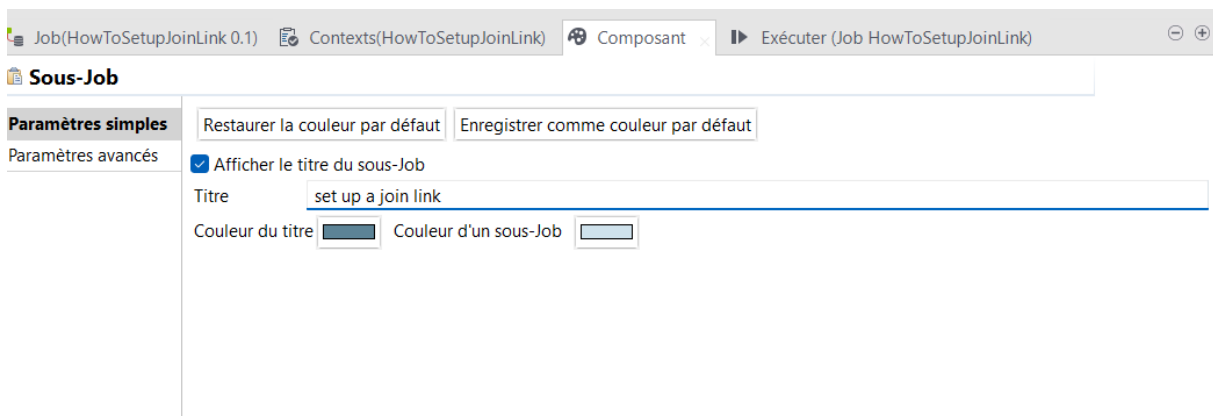


Voici le fichier customers+states.csv créé après l'exécution de la job:



id	CustomerName	CustomerAddress	LabelState
1	Griffith Paving	talend@apres	Connecticut
2	Bill's Dive Shop	511 Maple Ave	Ohio
3	Childress Chl	662 Lyons Circ	Alabama
4	Facelift Kitch	e unknown	
5	Terrinni & Son	770 Exmoor Rd.	
6	Kermit the Pet	1860 Parkside	Nevada
7	Tub's Furnitur	807 Old Trail R	Iowa
8	Toggle & Myer	618 Sheriden	Florida
9	Childress Chl	788 Tennyson	Idaho
10	Eile Hynosis	2032 Northbr	Alabama
11	Lennox Air Pol	4522 N. Greenview	Apt. 18
12	Keyth Contrac	1547 Knobwo	Missouri
13	Park District	C 2678 Sherida	Virginia
14	Nirabi Auto Se	1915 Lewis Ln	Delaware
15	Darcy Frame	1633 McGove	Massachusetts
16	Glenwood Cre	511 Maple Ave	Virginia
17	Gourmet the F	788 Tennyson	Alabama
18	Acturial Enterj	3385 Universit	North Dakota
19	Salt & Pepper	965 Marion Pl	Utah
20	Rythmics Ltd.	1875 Roger W	Michigan
21	Parkway Auto	1859 Green B	Nebraska
22	Futoons Carto	1486 Old Deer	Mississippi
23	Glenn Oaks O	1882 St. Johns	Georgia
24	Garfield Appli	3150 Skokie V	North Carolina
25	New Dehli Aut	1957 Huntingt	West Virginia
26	Kermit the Pet	12 Skokie Vall	Vermont
27			

Étape 5: Documenter le job



Job(HowToSetupJoinLink 0.1) Contexts(HowToSetupJoinLink) Composant Exécuter (Job HowToSetupJoinLink)

Sous-Job

Paramètres simples

Restaurer la couleur par défaut Enregistrer comme couleur par défaut

Paramètres avancés

☒ Afficher le titre du sous-Job

Titre

Couleur du titre Couleur d'un sous-Job

Conclusion

Ce TP met en lumière l'efficacité du composant tMap pour les transformations complexes. La jointure réalisée illustre une pratique courante dans les pipelines ETL, où des données provenant de différentes sources sont consolidées pour répondre à des besoins analytiques.

Conclusion générale

Ces deux TP fournissent une vue d'ensemble sur les étapes clés de l'intégration de données avec Talend. Le premier TP introduit les bases de la gestion des métadonnées, tandis que le second explore la transformation et l'agrégation des données. La maîtrise de ces techniques est cruciale pour la construction de workflows ETL robustes.

