

Talend Open Studio for Data Integration

Prendre en main le composant tMap - Filtre et rejet

Dans ce tutorial, vous utiliserez les métadonnées créées dans le tutorial "Créer une métadonnée de connexion à un fichier délimité" pour découvrir les particularités du composant tMap. Il sera utilisé pour sélectionner les données entrantes et créer trois sorties différentes :

- les clients de l'Ohio ;
- les clients des autres Etats ;
- les erreurs issues de la jointure entre les deux fichiers d'entrée.

Pour une meilleure consultation, ce tutorial est présenté en quatre parties :

- Prendre en main le composant tMap
- Prendre en main le composant tMap - Fonctionnalités essentielles
- Prendre en main le composant tMap - Créer une jointure
- Prendre en main le composant tMap - Filtre et rejet

Pré-requis :

Pour suivre ce tutorial, vous avez besoin d'extraire et d'installer les fichiers customer.csv et state.txt



Ajouter les filtres

Vous pouvez remarquer que l'adresse du client id=4 n'est pas renseignée (undefined). Il faut donc ajouter une sortie pour regrouper ces données erronées ou incomplètes.

De plus, l'objectif initial est de filtrer les clients de l'Ohio en sortie. Donc nous allons ajouter une sortie de tri des données.

Dans le Job Designer :

Double-cliquez sur le tMap pour ouvrir l'éditeur du tMap.

The screenshot shows the Talend Open Studio Job Designer interface. The job is named "Job TakeadvantageOfMapcomponent 0.1". The job design is as follows:

- A "states" component (Lookup) feeds into a "customer (Main)" component.
- The "customer (Main)" component feeds into a "tMap 1" component.
- The "tMap 1" component feeds into a "demo1 (Main)" component.
- The "demo1 (Main)" component feeds into a "tLogRow 1" component.

The "customer (Main)" component has the following data:

id	Customer...
4	Facelift Kitchen
unknown	unknown
idState	0
id2	15
ReqTime	22/08/2002 09:...

The "demo1 (Main)" component has the following data:

id	sumID	Customer...	LabelState
3	31	CHILDRESS CHI...	Alabama
idState	1	662 Lyons Circle	

The "tLogRow 1" component has the following data:

id	LabelState
3	Alabama

The bottom panel shows the "Job TakeadvantageOfMapcomponent" execution controls, including "Resume", "Kill", "Save Job before execution", "Clear before run", and "Exec time" options. The "Stats & Traces" section is also visible, with "Statistics" and "Traces" checkboxes.

Dans l'éditeur du tMap :

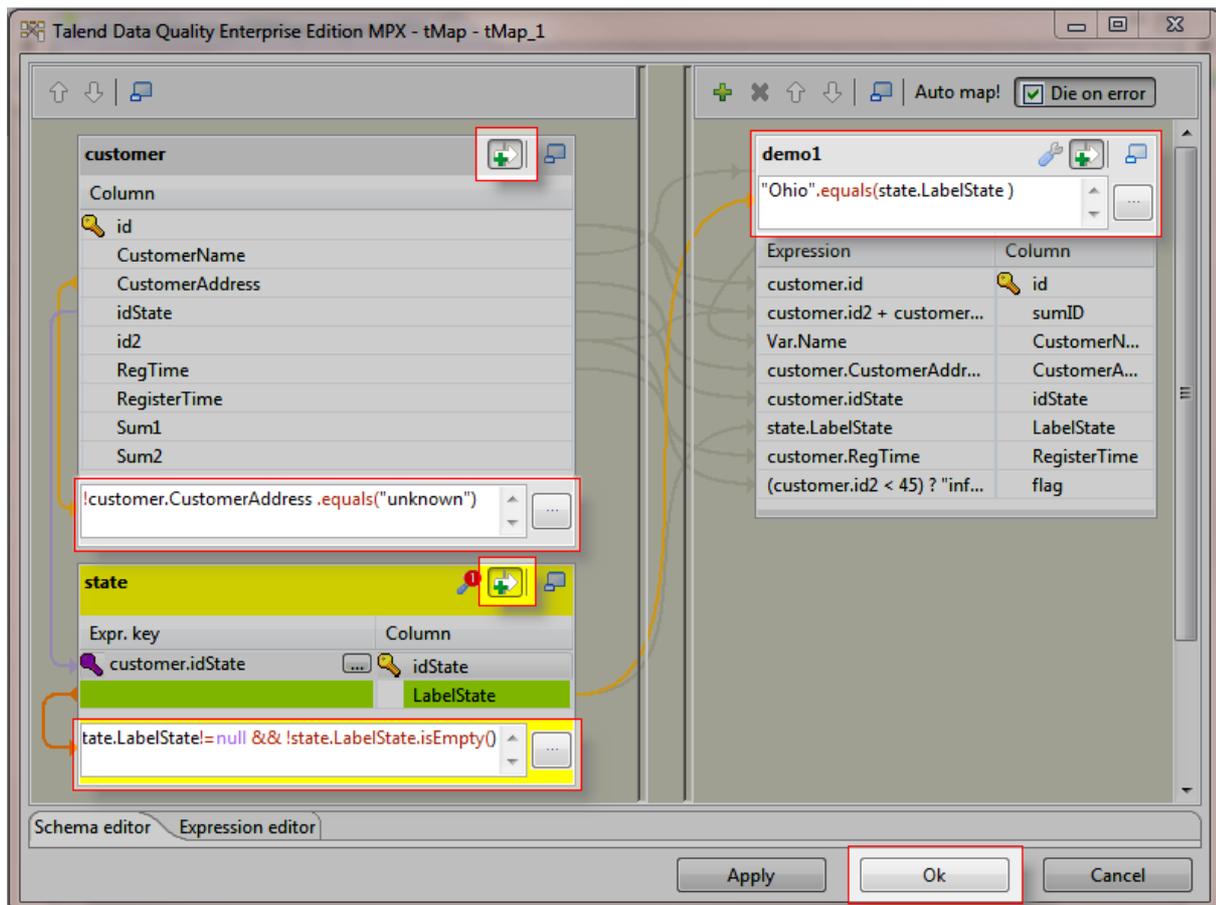
Cliquez sur la flèche blanche en haut de chaque table pour ajouter le champ de filtre.

Dans la table customer, saisissez !, appuyez sur Ctrl+Espace, double-cliquez sur customer.CustomerAddress dans la liste d'autocomplétion et saisissez l'expression Java `!.equals("unknown")`.

Dans la table state, saisissez `state.LabelState != null && !state.LabelState.isEmpty()` pour éviter les champs nuls et les champs vides.

Dans la table demo1, saisissez l'expression Java `:"Ohio".equals(state.LabelState)`.

Cliquez sur OK.



2

Exécuter le Job

Dans le Job Designer :

Avant d'exécuter votre Job, enregistrez-le via Ctrl+S.

Cliquez sur la vue Run.

Dans la vue Run :

Cochez la case Statistics.

Cliquez sur le bouton Run pour exécuter le Job en mode Statistics.

The screenshot displays the Job Designer interface for a job named "Job TakeadvantageOfMapcomponent 0.1". The main workspace shows a flow diagram with the following components and data:

- customers**: 5000 rows in 0.93; 5370.57 rows/s; customer (Main)
- tMap_1**: 110 rows in 0.93; 118.15 rows/s; demo1 (Main)
- states**: 51 rows in 0; 12750 rows/s; state (Lookup)
- tLogRow_1**: Output component

The "Run" dialog box is open, showing the "Execution" tab. The "Stats & Traces" section has "Statistics" checked. The execution log shows the following output:

```
Drive|35|Ohio|04/05/1990 14:22|interior  
[statistics] disconnected  
Job TakeadvantageOfMapcomponent ended at 00:46 31/12/2014  
[exit code=0]
```

Ce Job est maintenant terminé !

Seuls les clients de l'Ohio sont transmis en sortie conformément au filtre défini dans l'éditeur du tMap.

3

Ajouter la sortie contenant les rejets

Vous pouvez remarquer que 110 clients sont basés dans l'Ohio dans le fichier.

Dans le Job Designer :

Double-cliquez sur le tMap pour ouvrir l'éditeur du tMap.

The screenshot displays the Talend Job Designer interface for a job named "Job TakeadvantageOfMapcomponent 0.1". The job flow in the Designer view consists of the following components and data:

- customers**: 5000 rows in 0.93, 5370.57 rows/s, customer (Main)
- tMap_1**: 110 rows in 0.93, 118.15 rows/s, demo1 (Main). This component is highlighted with a red box.
- tLogRow_1**: Receives data from tMap_1.
- states**: 51 rows in 0, 12750 rows/s, state (Lookup)

The bottom panel shows the execution details for "Job TakeadvantageOfMapcomponent":

- Context**: Target execution is set to "1".
- Execution**: Run button is active. Options include "Save Job before execution" (unchecked), "Clear before run" (checked), and "Exec time" (unchecked).
- Stats & Traces**: "Statistics" is checked, and "Traces" is unchecked.
- Log Output**:

```
Drive|35|Ohio|04/05/1990 14:22|interior
[statistics] disconnected
Job TakeadvantageOfMapcomponent ended at 00:46 31/12/2016
[exit code=0]
```

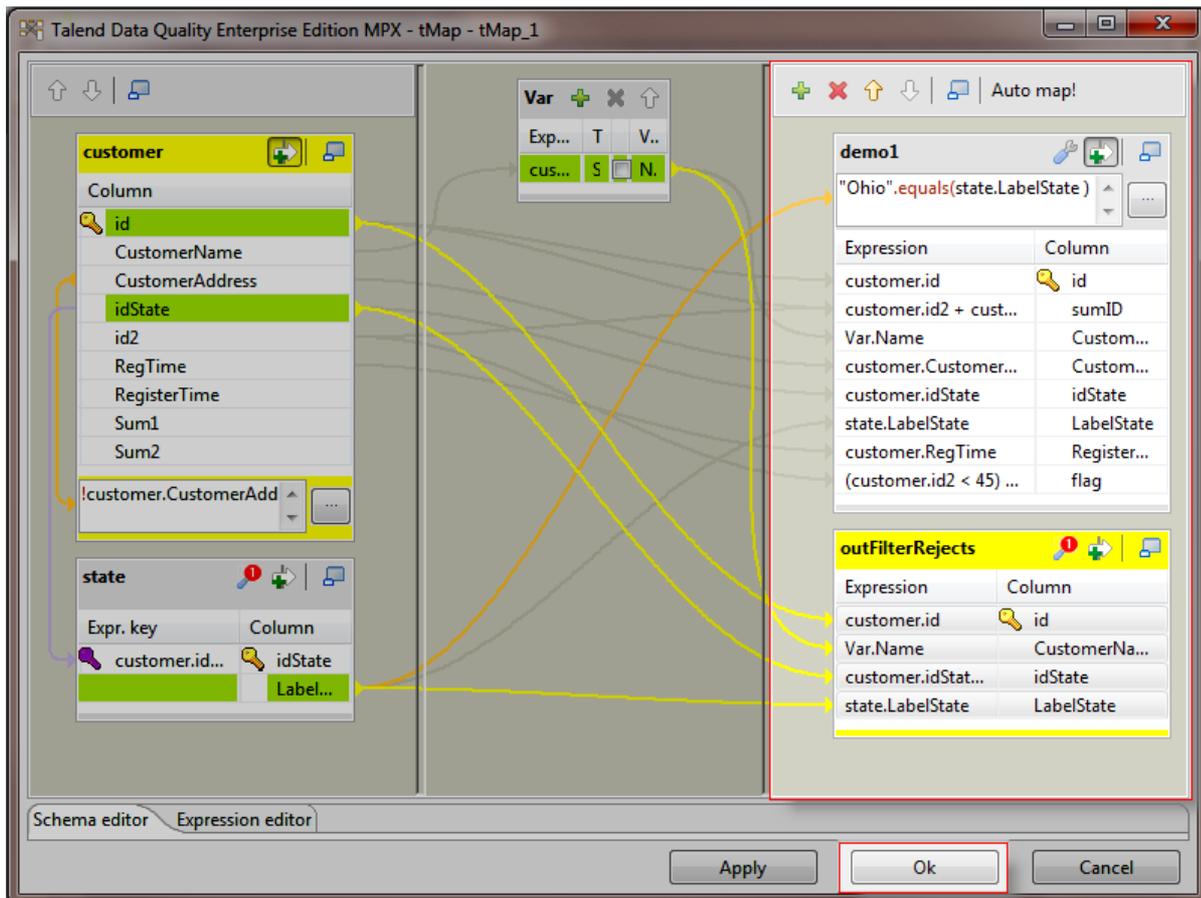
Dans l'éditeur du tMap :

Cliquez sur (+) dans la barre d'outils au dessus de la table demo1 pour ajouter une nouvelle table en sortie. Dans l'assistant Add a new output table, saisissez le nom que vous souhaitez donner à la table (outFilterRejects) et cliquez sur OK.

Dans la table outFilterRejects, cliquez sur le bouton avec la flèche orange pour collecter les données rejetées par le filtre de la sortie demo1 (Option « Catch output reject » à « true »).

Dans la table demo1, appuyez sur Ctrl tout en sélectionnant les colonnes customer.id, Var.Name, customer.idState et state.LabelState et faites-les glisser dans la table outFilterRejects.

Cliquez sur OK.

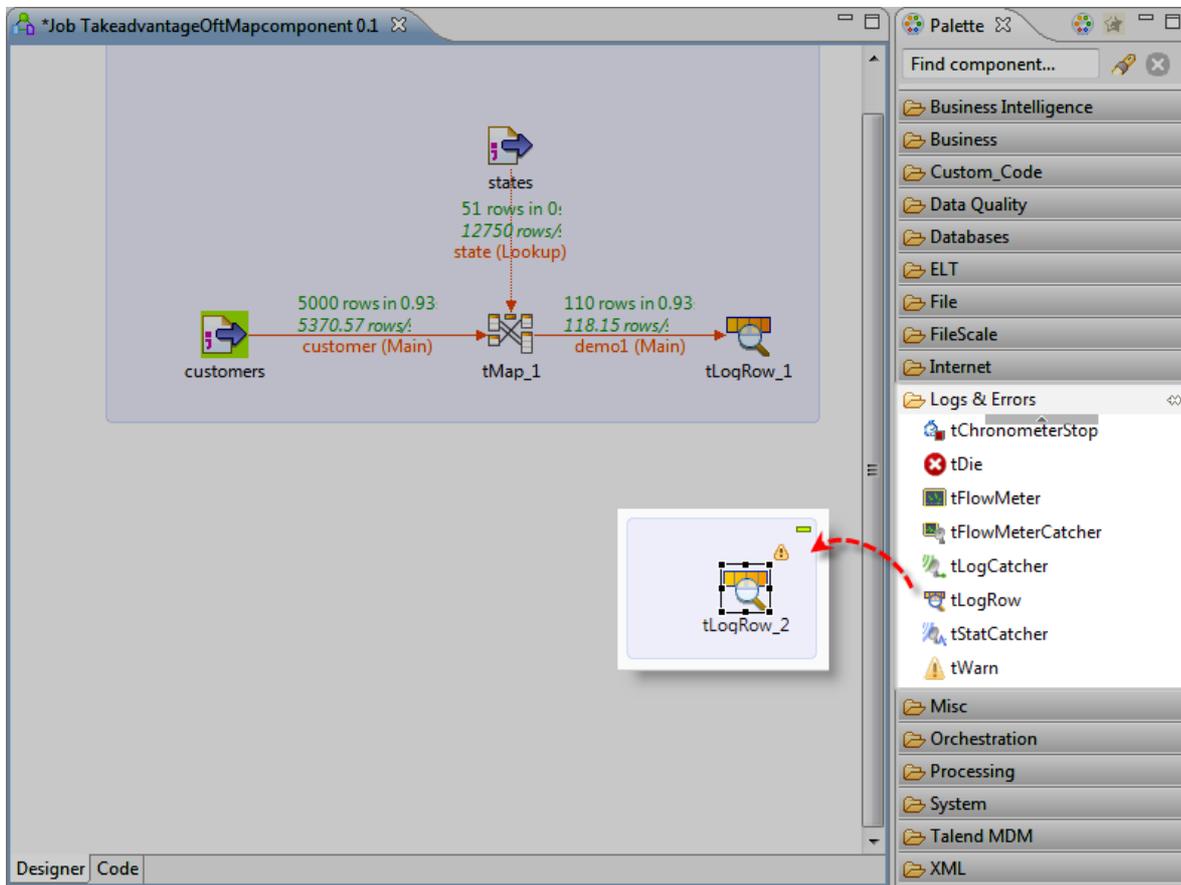


4 Déposer le composant de sortie des rejets et le relier au composant de transformation

Dans la Palette :

Pour ajouter le composant de sortie, cliquez sur la famille Logs & Errors.

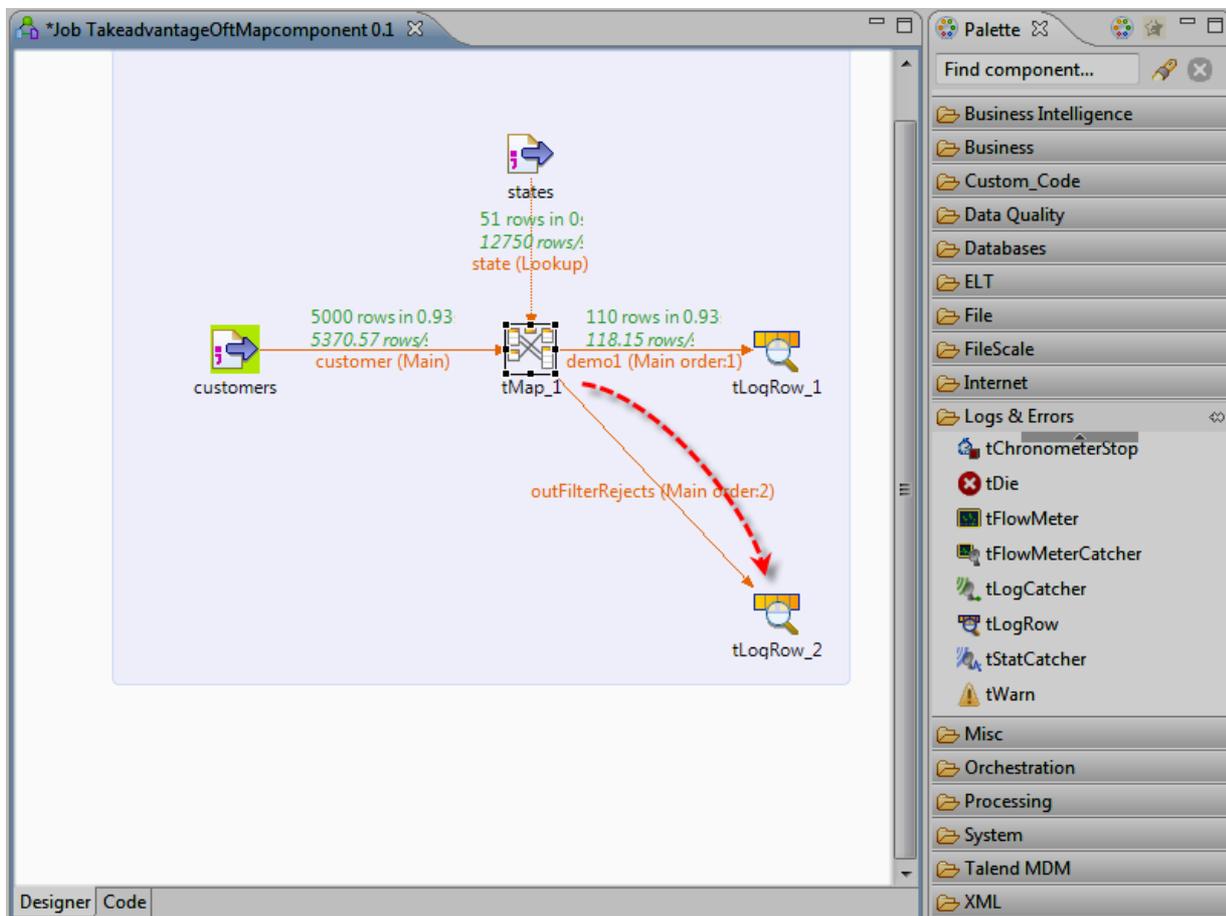
Cliquez sur le composant tLogRow et dépose-le dans le Job Designer.



Dans le Job Designer :

Cliquez-droit sur le composant tMap.

Dans le menu, sélectionnez Row > outFilterRejects et cliquez sur le tLogRow.



5

Exécuter le Job

Dans le Job Designer :

Avant d'exécuter votre Job, enregistrez-le via Ctrl+S.

Appuyez sur F6 pour lancer l'exécution du Job.

La vue Run s'affiche en bas de Talend Open Studio et la console retrace l'exécution du Job.

The screenshot displays the Talend Open Studio interface during a job execution. The top window, titled 'Job TakeadvantageOfMapcomponent 0.1', shows a job diagram with the following components and data flow statistics:

- customers**: 5000 rows in 30.64s, 163.17 rows/s
- tMap_1**: 110 rows in 30.64s, 3.59 rows/s
- tLogRow_1**: 4889 rows in 30.64s, 159.55 rows/s
- states**: 51 rows in 0s, 12750 rows/s

The bottom window, titled 'Run (Job TakeadvantageOfMapcomponent)', shows the execution console with the following log entry:

```
5000|LIBRA & SUNS|27|Nebraska  
[statistics] disconnected  
Job TakeadvantageOfMapcomponent ended at 01:02 31/12/2016  
/exit code=0
```

Le job fonctionne mais il reste encore des données incomplètes.

Il s'agit donc maintenant de les isoler dans une nouvelle sortie.

6

Ajouter la sortie contenant les données erronées

Dans le Job Designer :

Si vous exécutez le Job à nouveau en mode « Debugage de Traces », vous pouvez observer que le champ idState du client id=11 n'a pas de nom d'état correspondant.

Double-cliquez sur le composant tMap pour ouvrir l'éditeur du tMap.

The screenshot shows the Job Designer interface for a job named "TakeadvantageOfMapcomponent 0.1". The main workspace displays a flow diagram with the following components and data:

- customers**: 43 rows - 48.05, 0.89 rows/s. Current row: 43.
- states**: 51 rows in 0.02, 2550 rows/s. state (lookup).
- tMap_1**: The central transformation component.
- customer (Main)**: 43 rows - 48.05, 0.89 rows/s. Current row: 43.
- demo1 (Main order:1)**: 1 rows - 1.83, 0.55 rows/s. Current row: 1.
- outFilterRejects (Main order:2)**: 42 rows. Current row: 42.

The data tables shown in the screenshot are as follows:

customer (Main)	
id	44
CustomerName	Parkway Auto Body
CustomerAddress	726 White Oaks Ave.
idState	14
id2	47
ReqTime	11/05/1993 14:35
RegisterTime	2000-06-15 21:13:20.000
Sum1	94320.0
Sum2	69425.89

demo1 (Main order:1)	
id	41
sumID	0.85
CustomerName	DIVE SHOP
CustomerAddress	511 Maple Ave. Apt. 1B
idState	35
LabelState	Ohio
RegisterTime	19/11/2004 15:48
flag	inferior

outFilterRejects (Main order:2)	
id	44
CustomerName	PARKWAY AUTO BODY
idState	14
LabelState	

Dans l'éditeur du tMap :

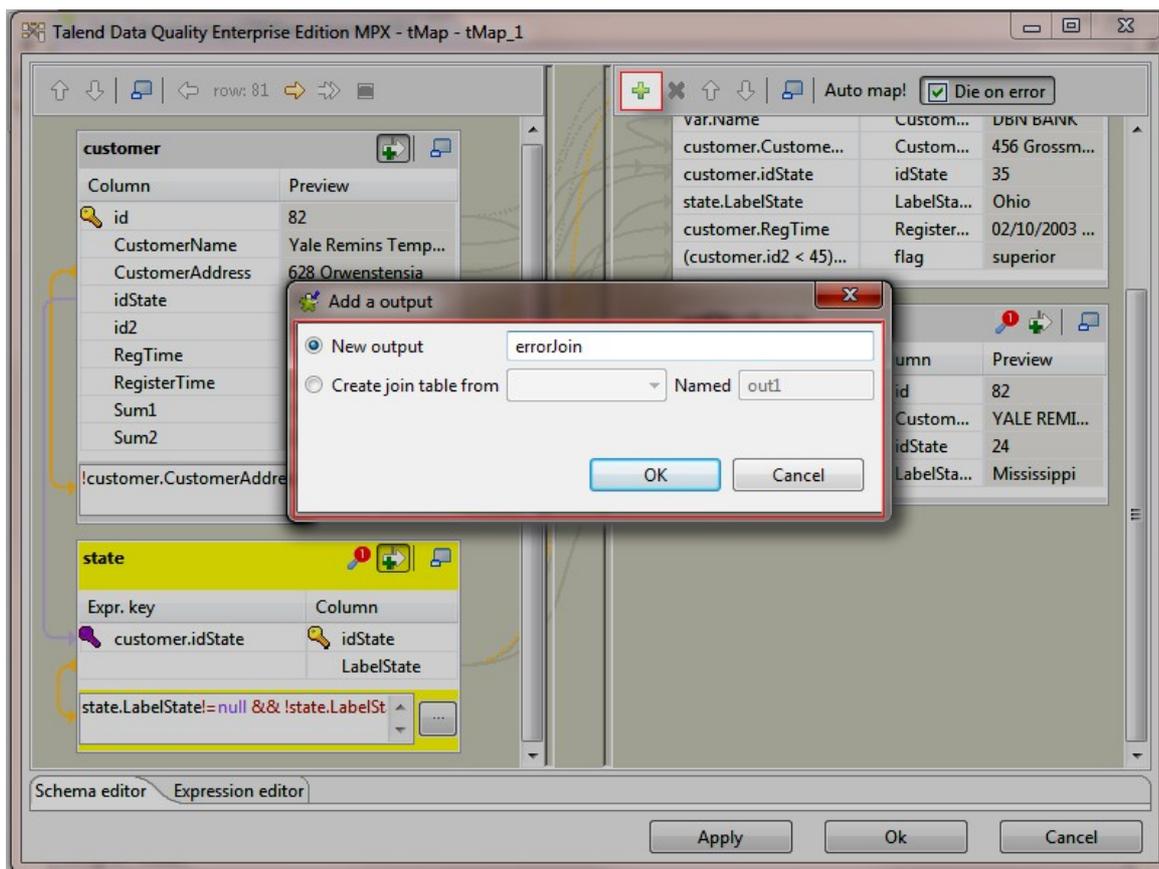
Cliquez sur (+) dans la barre d'outils au dessus de la table demo1 pour ajouter une nouvelle table de sortie. Dans l'assistant Add a new output table, saisissez le nom que vous souhaitez donner à la table (errorJoin) et cliquez sur OK.

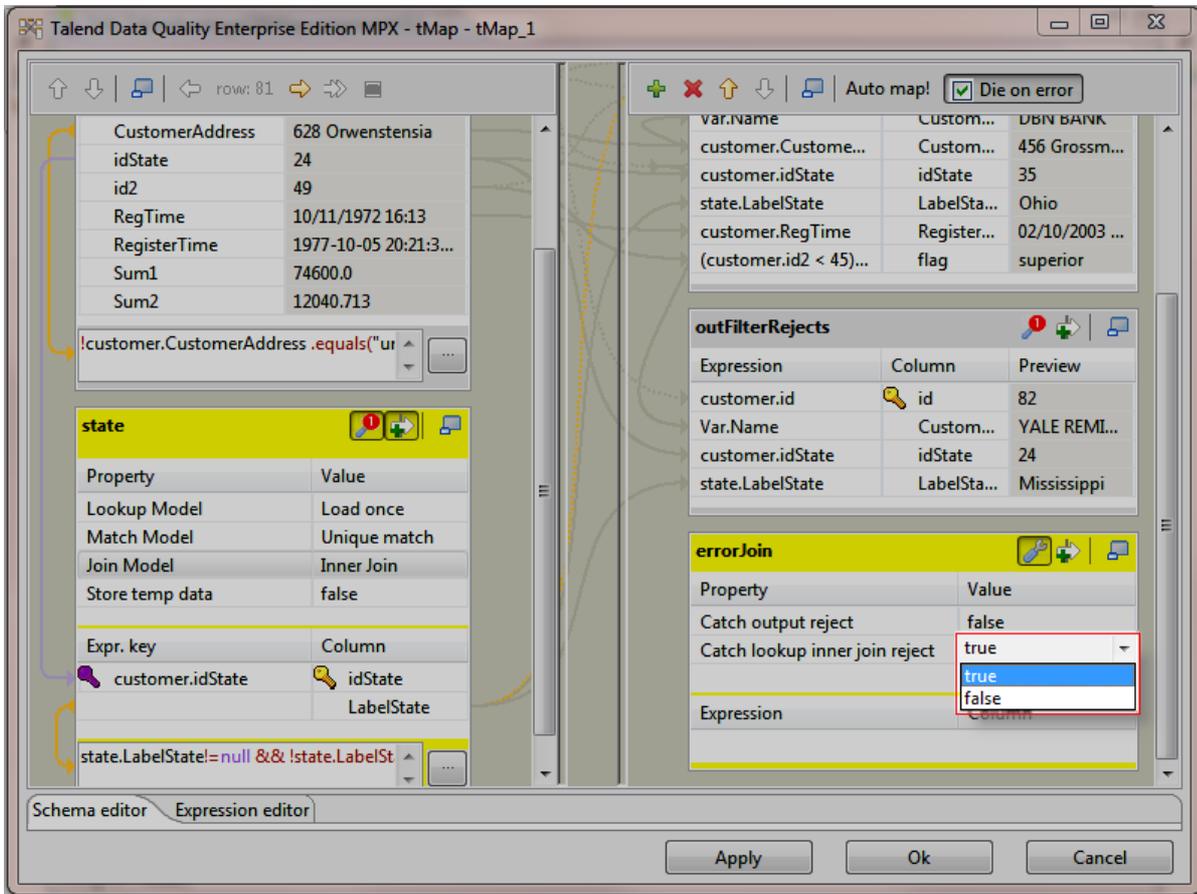
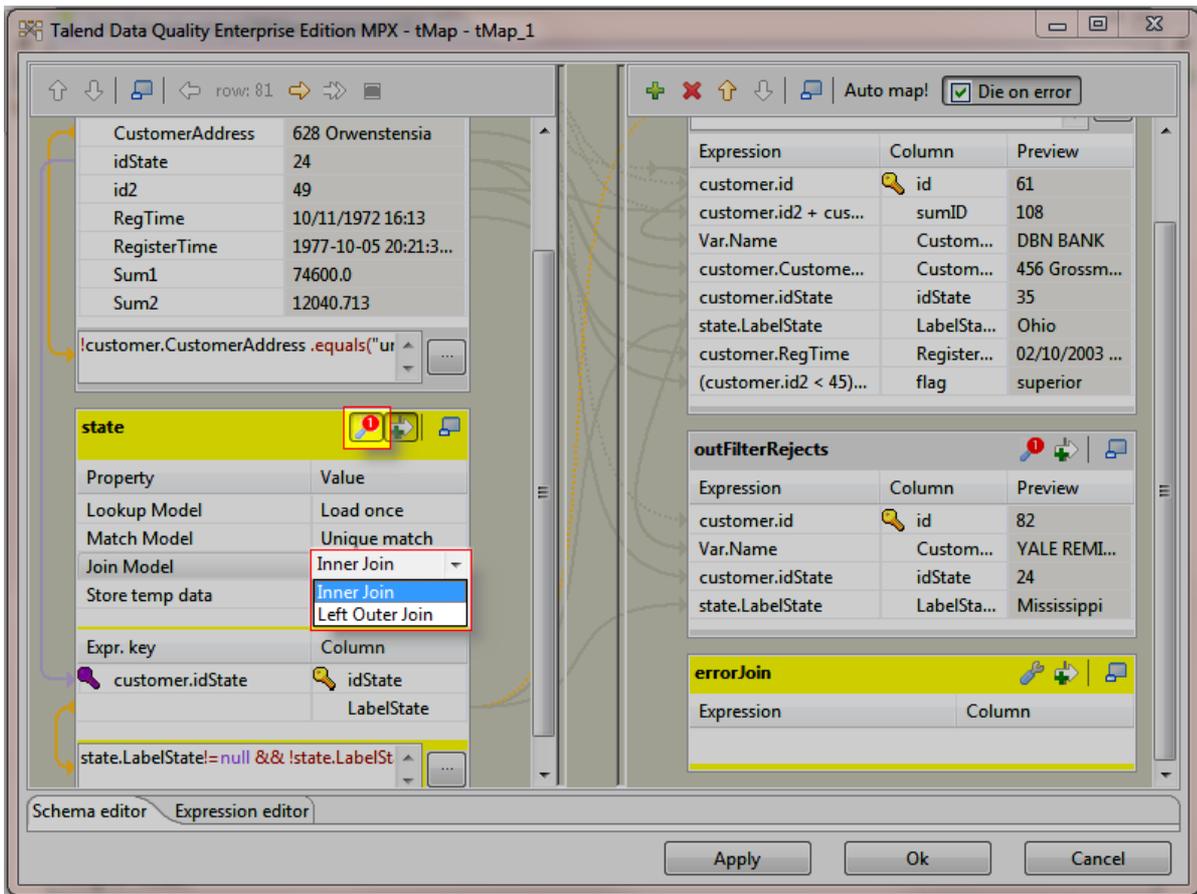
Cliquez sur le bouton Inner Join de la table state.

Cliquez sur la flèche violette de la table errorJoin pour collecter les données rejetées par la jointure (Option « Catch output reject » à « true »).

Dans la table customer, appuyez sur Ctrl tout en sélectionnant les colonnes id, CustomerName, CustomerAddress et idState et faites-les glisser dans la table errorJoin.

Cliquez sur OK.





Talend Data Quality Enterprise Edition MPX - tMap - tMap_1

row: 81

Auto map! Die on error

Column	Preview
id	82
CustomerName	Yale Re...
CustomerAddress	628 Orw...
idState	24
id2	49
RegTime	10/11/1...
RegisterTime	1977-10...
Sum1	74600.0
Sum2	12040.713

customer.CustomerAddress.equals(

Expr. key	Column
customer.idState	idState
	LabelState

Expression	Column	Preview
customer.RegTime	Register...	02/10/2003 ...
(customer.id2 < 45)...	flag	superior

outFilterRejects

Expression	Column	Preview
customer.id	id	82
Var.Name	Custom...	YALE REMI...
customer.idState	idState	24
state.LabelState	LabelSta...	Mississippi

errorJoin

Property	Value
Catch output reject	false
Catch lookup inner join reject	true

Expression	Column
customer.id	id
customer.CustomerName	CustomerName
customer.CustomerAddress	CustomerAdd...
customer.idState	idState

Schema editor Expression editor

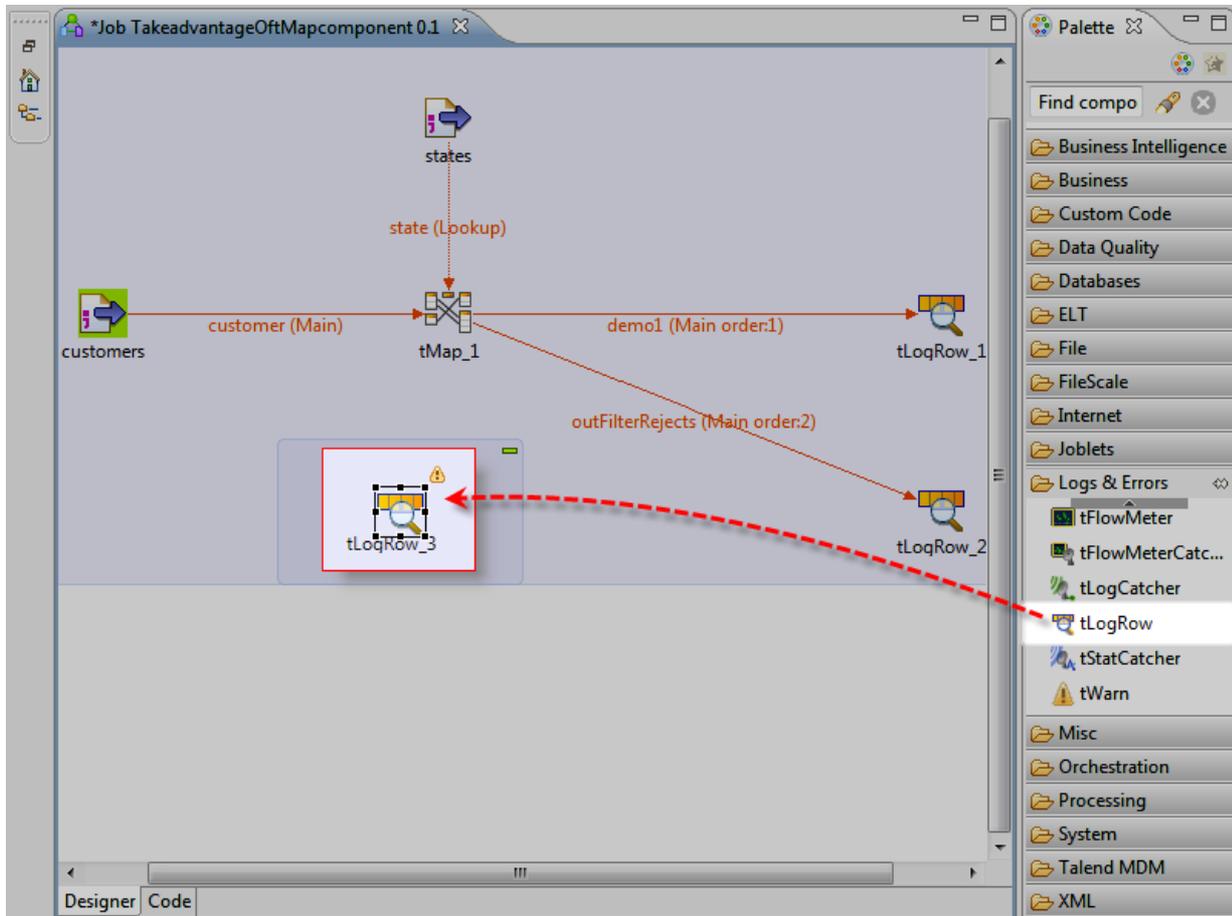
Apply Ok Cancel

7 Déposer le composant de sortie des erreurs et le relier au composant de transformation

Dans la Palette :

Pour ajouter le composant de sortie, cliquez sur la famille Logs & Errors.

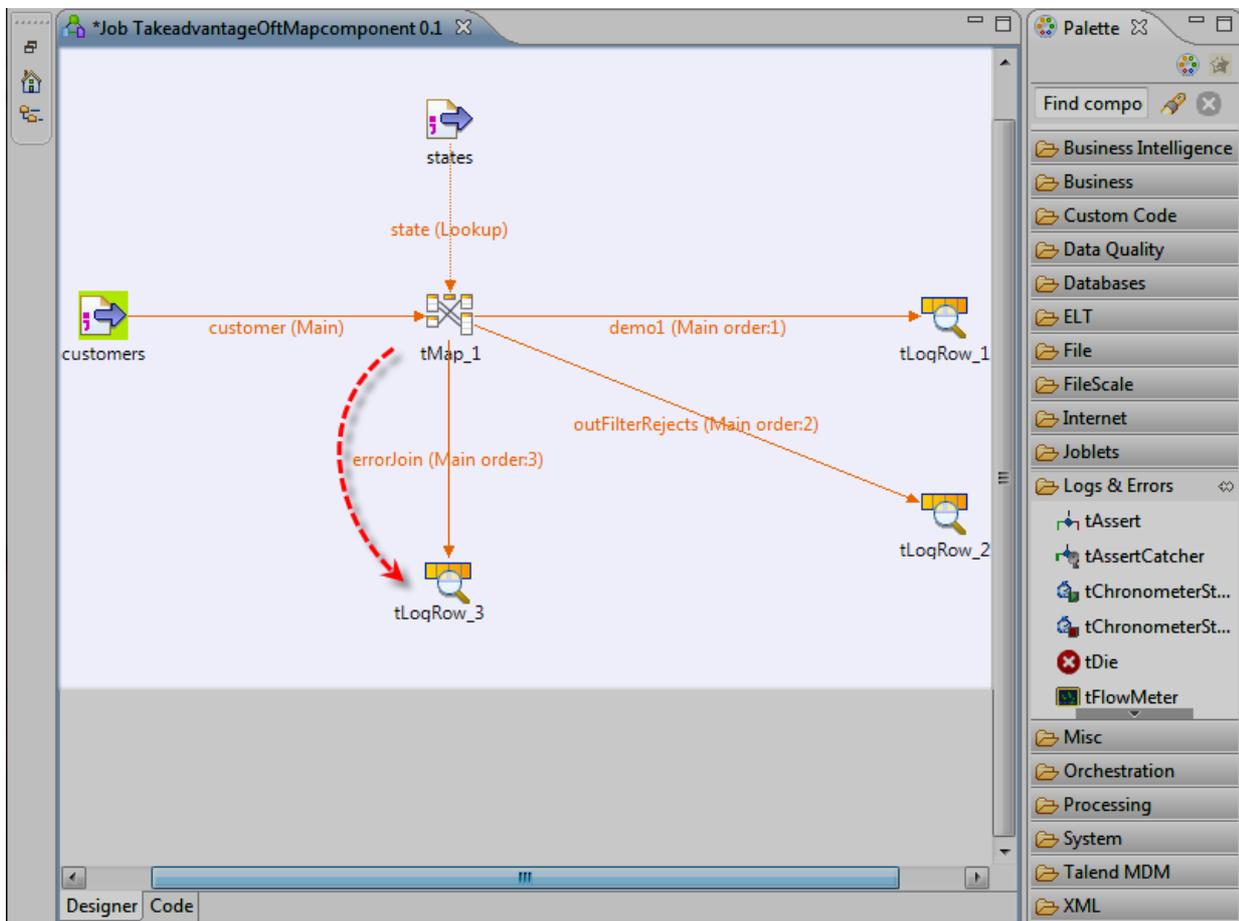
Cliquez sur le composant tLogRow et déposez-le dans le Job Designer.



Dans le Job Designer :

Cliquez-droit sur le composant tMap.

Dans le menu contextuel, sélectionnez Row >errorJoin et cliquez sur le composant tLogRow.





Exécuter le Job

Dans le Job Designer :

Avant d'exécuter votre Job, enregistrez-le via Ctrl+S.

Cliquez sur la vue Run.

Dans la vue Run :

Cochez la case Statistics.

Cliquez sur le bouton Run pour lancer l'exécution du Job en mode Statistics.

The screenshot displays the Job Designer interface for a job named "Job TakeadvantageOfMapcomponent 0.1". The main workspace shows a flow diagram with the following components and data:

- customers**: 5000 rows in 39.12, 127.8 rows/s
- states**: 51 rows in 0.16, 326.92 rows/s, state (Lookup)
- tMap_1**: Receives input from customers and states.
- errorJoin (Main order:3)**: 1 rows in 39.12, 0.03 rows/s
- demo1 (Main order:1)**: 110 rows in 39.12, 2.81 rows/s
- outFilterRejects (Main order:2)**: 4888 rows in 39.12, 124.93 rows/s
- tLogRow_1**: Receives input from demo1 and outFilterRejects.

Below the workspace is the "Run (Job TakeadvantageOfMapcomponent)" window. It features a "Basic Run" tab with "Run", "Kill", and "Clear" buttons. The console output shows:

```
5000|LIBRA & SONS|27|Nebraska  
[statistics] disconnected  
Job TakeadvantageOfMapcomponent ended at 22:50  
09/02/2011. [exit code=0]
```

The "Line limit" is set to 100.

Le Job TakeadvantageOfMapcomponent est maintenant terminé !

Il crée une jointure entre deux fichiers d'entrée et génère trois sorties différentes :

- une pour les clients de l'Ohio,
- une pour les clients des autres Etats,
- une pour les données erronées ou incomplètes.

Il ne reste plus qu'à le documenter !

Dans le Job Designer :

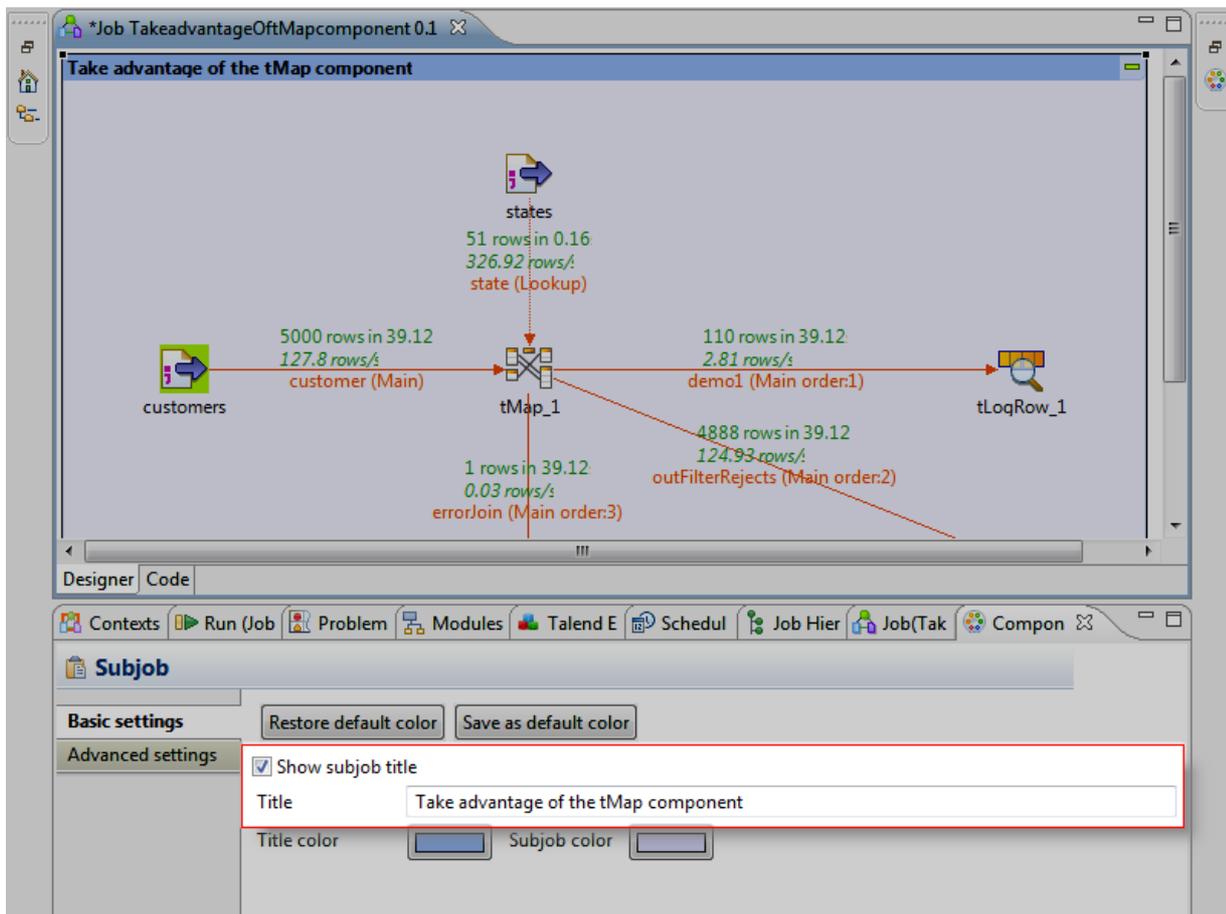
Pour documenter votre Job, donnez-lui un titre.

Pour cela, cliquez sur la zone bleue entourant votre Job.

Cliquez sur la vue Component.

Pour lui ajouter un titre, cochez la case Show subjob title et dans le champ Title, saisissez le titre correspondant Take advantage of the tMap component (Apprendre à utiliser le composant tMap, en français).

Enregistrez de nouveau le Job.



Ce tutoriel est maintenant terminé.

Le Job fonctionne et est documenté.