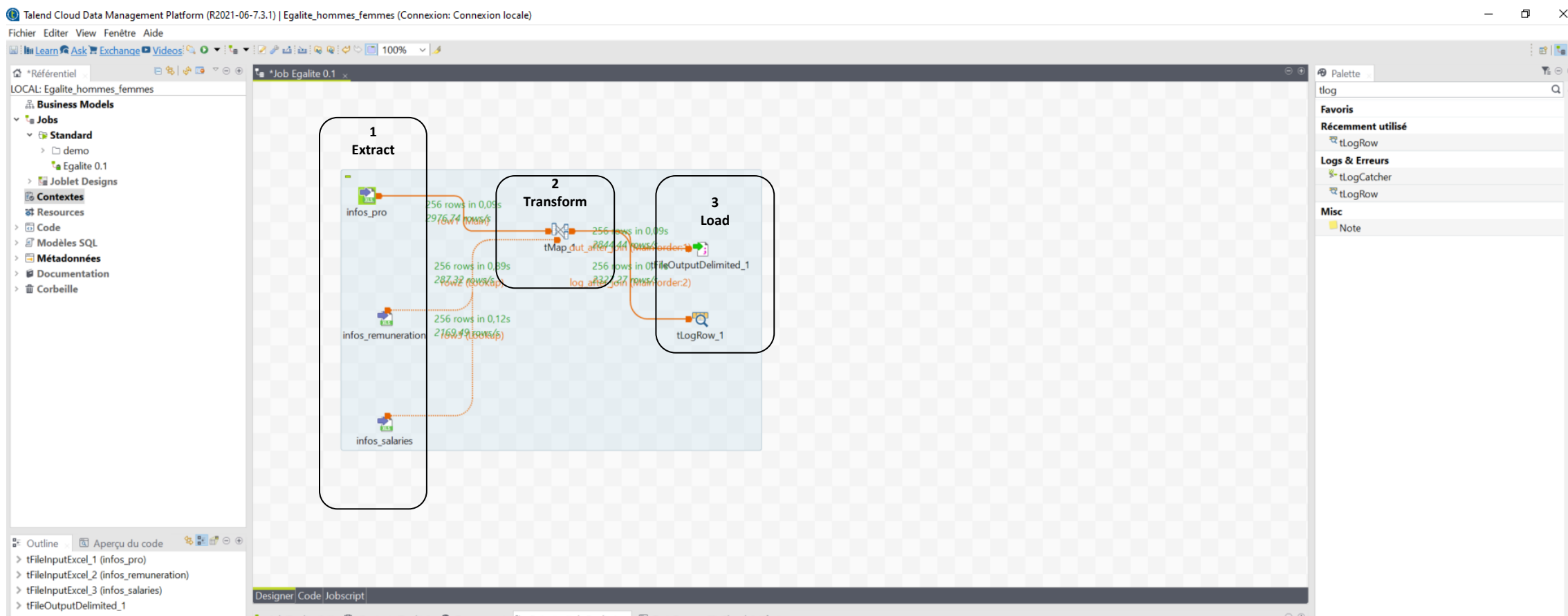


Utilisation de l'ETL TALEND® pour la création d'un fichier .CSV avec données anonymisées



1 Extract

Nos trois fichiers source présentent des données à caractère personnel sur les salariés d'une entreprise.

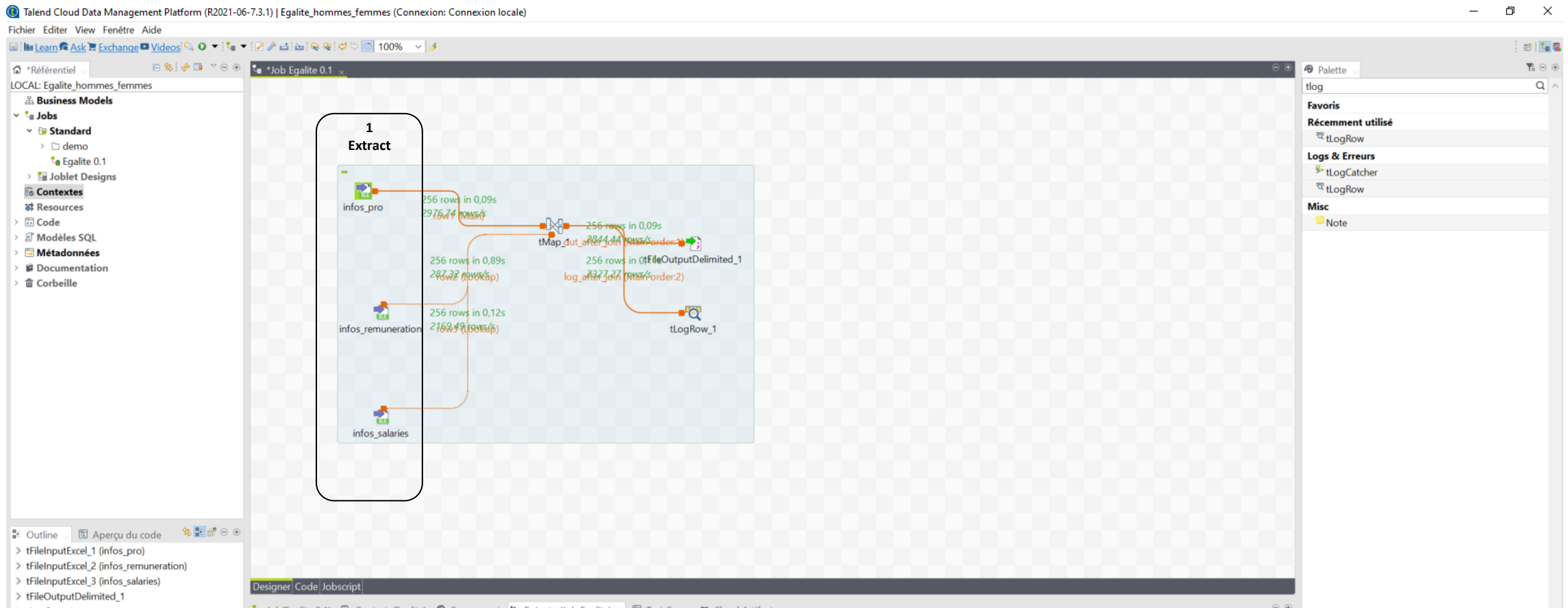
Objectif : créer un fichier CSV avec les données issues des trois fichiers source, dans le respect du RGPD. Ce fichier sera utilisé pour la réalisation d'un diagnostic égalité professionnelle hommes-femmes (embauche, qualification, promotion, conditions de travail , rémunération...)

=> suppression de certaines variables non utiles (nom, prénom, état civil, nombre d'enfants...)

=> discrétisation de certaines variables numériques (âge, ancienneté...)

=> création de nouvelles variables (salaire total...)

Utilisation de l'ETL TALEND® pour la création d'un fichier .CSV avec données anonymisées



1.1 Extract : table_info_pro.xlsx

The screenshot displays the Talend Cloud Data Management Platform interface (R2021-06-7.3.1) for a job named "Gender_equality (Connexion: Talend_local)". The job is titled "Job Gender_equality_indicators 0.1" and is located in the "Business Models" section under "Jobs" > "Standard".

The job design is visible, showing a flow from "infos_pro" (256 rows in 0.1s) to "infos_remuneration" (256 rows in 0.14s) and "infos_salaries" (256 rows in 0.14s). The "infos_pro" component is highlighted with a red box, and a red arrow points to its configuration panel.

The configuration panel for "infos_pro(tFileInputExcel_1)" is shown, detailing the following parameters:

- Type de propriété: Référentiel
- ExcelInfo: infos_pro
- Paramètres avancés:
 - ☒ Lire un fichier au format excel2007 (xlsx)
 - Nom de fichier/Flux: "C:/Users/Celine/workspace/GENDER_EQUALITY/Table info_pro.xlsx"
 - Password: *****
- Paramètres dynamiques: (empty)
- View: (empty)
- Documentation: (empty)
- Règles de validation:
 - ☒ Toutes les feuilles
 - En-tête: 1
 - Pied de page: 0
 - Limite: (empty)
 - ☐ Affecte chaque feuille (en-tête et pied de page)
 - ☐ Arrêter en cas d'erreur
 - Première colonne: 1
 - Dernière colonne: (empty)
 - Schéma: Référentiel
 - ExcelInfo: infos_pro - metadata
 - Modifier le schéma: (empty)

The right sidebar shows a "Palette" of components, including "Favoris", "Récemment utilisé", "Applications Métier", "Bases de données", "Big Data", "Business Intelligence", "Business", "Cloud", "Code Utilisateur", "Data Privacy", "Databases NoSQL", "Databases", "Divers", "DotNET", "ELT", "Combined SQL", "Connections", "Connexions", "Map", "SQLTemplate", "ESB", "Fichier", "Gestion", "Lecture", "NamedPipe", "Écriture", "Internet", "Logs & Erreurs", "tAssert", "tAssertCatcher", "tChronometerStart", "tChronometerStop", "tDie", "tFlowMeter", "tFlowMeterCatcher", "tLogCatcher", "tLogRow", "tStatCatcher", "tWarn", and "Orchestration".

1.2 Extract : table_remuneration.xlsx

The screenshot displays the Talend Cloud Data Management Platform interface for a job named "Gender_equality (Connexion: Talend_local)". The job is configured to extract data from an Excel file named "table_remuneration.xlsx".

Job Design: The job design shows a sequence of components: "infos_pro" (256 rows in 0.1s), "infos_remuneration" (256 rows in 0.75s), "infos_salaries" (256 rows in 0.14s), "tMap_1" (256 rows in 0.1s), "tLogRow_1" (256 rows in 0.1s), and "tFileOutputDelimited_1" (256 rows in 0.1s). A red box highlights the "infos_remuneration" component, which is linked to the "tFileInputExcel_2" component in the properties panel.

Properties Panel: The properties panel for "infos_remuneration (tFileInputExcel_2)" is shown, detailing the extraction parameters:

- Type de propriété:** Référentiel
- Excel Infos:** infos_remuneration
- Paramètres simples:**
 - ☒ Lire un fichier au format excel2007 (xlsx)
 - Nom de fichier/Flux:** "C:/Users/Celine/workspace/GENDER_EQUALITY/Table remuneration.xlsx"
 - Password:** *****
 - ☒ Toutes les feuilles
 - En-tête:** 1
 - Pied de page:** 0
 - ☐ Affecte chaque feuille (en-tête et pied de page)
 - ☐ Arrêter en cas d'erreur
 - Première colonne:** 1
 - Dernière colonne:** (empty)
 - Schéma:** Référentiel
 - Excel Infos:** infos_remuneration - it
 - ☐ Modifier le schéma

1.3 Extract : table_salaries.xlsx

The screenshot displays the Talend Cloud Data Management Platform interface for a job named "Job Gender equality.indicators 0.1". The job is configured to extract data from an Excel file named "table_salaries.xlsx".

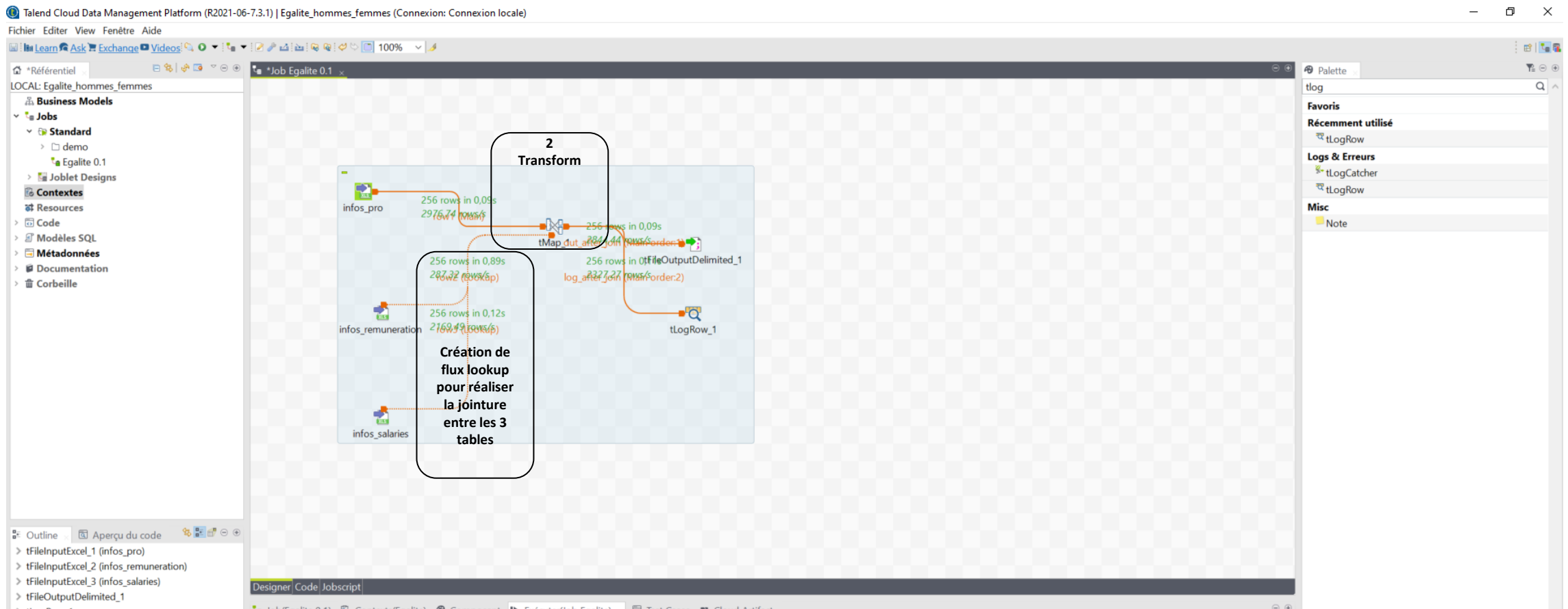
Job Design: The job design shows a sequence of components: "infos_pro" (256 rows in 0.1s), "tMap_1" (256 rows in 0.75s), "data_output" (256 rows in 0.1s), "tLogRow_1" (256 rows in 0.1s), and "tFileOutputDelimited_1" (256 rows in 0.1s). A component named "infos_salaries" is highlighted with a red box and a callout arrow pointing to its configuration panel.

Configuration Panel for "infos_salaries":

- Type de propriété:** Référentiel
- Paramètres simples:**
 - ☒ Lire un fichier au format excel2007 (xlsx)
 - Nom de fichier/Flux:** "C:/Users/Celine/workspace/GENDER_EQUALITY/Table salaries.xlsx"
 - Password:** [Redacted]
 - ☒ Toutes les feuilles
 - En-tête:** 1
 - Pied de page:** 0
 - Limite:** [Empty]
 - ☐ Affecte chaque feuille (en-tête et pied de page)
 - ☐ Arrêter en cas d'erreur
 - Première colonne:** 1
 - Dernière colonne:** [Empty]
 - Schéma:** Built-In
 - ☐ Modifier le schéma
- Paramètres avancés:** [Empty]
- Paramètres dynamiques:** [Empty]
- View:** [Empty]
- Documentation:** [Empty]
- Règles de validation:** [Empty]

2 Transform

tMap



TMap : Jointure interne entre les 3 tables

Inner Join on *id_salarie*

The screenshot displays the Talend Cloud Data Management Platform interface for configuring a TMap job. The main window shows three input tables (row1, row2, row3) and their columns. A callout box highlights the 'Inner Join on id_salarie' configuration. The 'Mapping auto' panel on the right shows the resulting columns for the join. The 'data_after_join' table at the bottom lists the columns and their data types.

Colonne	Type	N.	Modèle de date (Ctrl+Espace)	Length	Precision	Défaut	Commentaire
id_salarie	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	13	0		
Anciennete_an	Float	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		0		
Distance_domicile_Travail	Integer	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	2	0		
Service	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	15	0		
Work_accident	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	3	0		
Niveau_de_satisfaction	Integer	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	3	0		

Colonne	Type	N.	Modèle de date (Ctrl+Espace)	Length	Precision	Défaut	Commentaire
Service	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	15	0		
Sexe	Character	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	1	0		
Age_binned	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		0		
Contrat	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	3	0		
Duree_hebdo	Integer	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	2	0		
Seniority_binned	String	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		0		
Salaire_total	Float	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	9			
Augmentation	Integer	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	1	0		
Promotion	Integer	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	1	0		

TMap : constructeur d'expression

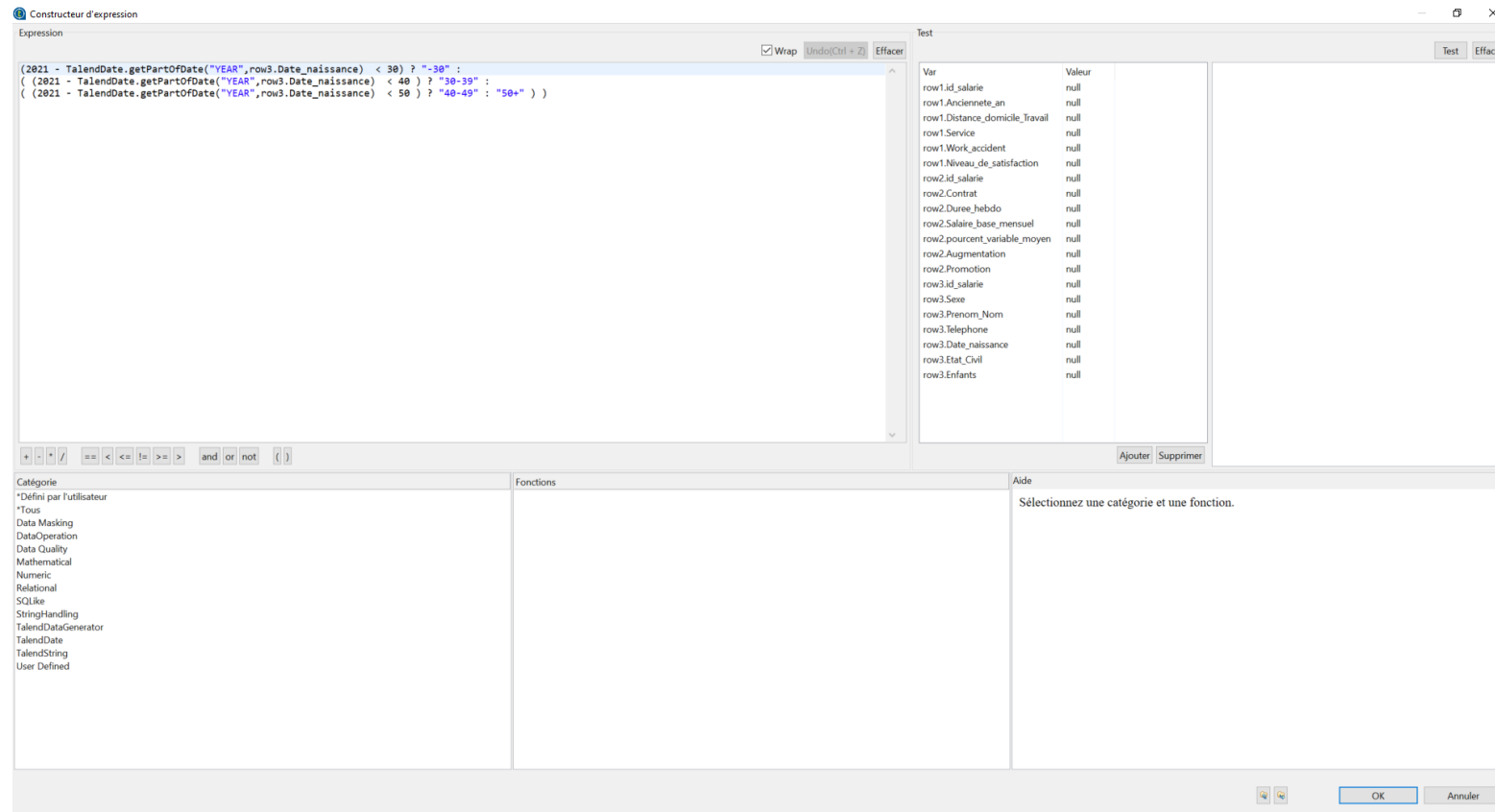
The screenshot displays the Talend Cloud Data Management Platform interface, specifically the TMap component. The main workspace is divided into three panes: 'row1', 'row2', and 'row3'. Each pane shows a list of columns and their corresponding data types. The 'row1' pane lists columns like 'id_salarie', 'Anciennete_an', 'Distance_domicile_Travail', 'Service', 'Work_accident', and 'Niveau_de_satisfaction'. The 'row2' and 'row3' panes show columns like 'id_salarie', 'Contrat', 'Duree_hebdo', 'Salaire_base_mensuel', 'pourcent_variable_moyen', 'Augmentation', and 'Promotion'. A central pane shows a mapping diagram with lines connecting columns between rows. On the right, a 'Mapping auto' pane shows the 'data_after_join' table with columns like 'Service', 'Sexe', 'Age_binned', 'Contrat', 'Duree_hebdo', 'Seniority_binned', 'Salaire_total', 'Augmentation', and 'Promotion'. Below the main workspace, there are two tables: 'row1' and 'data_after_join'. The 'row1' table has columns for 'Colonie', 'Type', 'N. Modèle de date (Ctrl+Espace)', 'Length', 'Precision', 'Défaut', and 'Commentaire'. The 'data_after_join' table has columns for 'Colonie', 'Type', 'N. Modèle de date (Ctrl+Espace)', 'Length', 'Precision', 'Défaut', and 'Commentaire'. The 'data_after_join' table lists columns like 'Service', 'Sexe', 'Age_binned', 'Contrat', 'Duree_hebdo', 'Seniority_binned', 'Salaire_total', 'Augmentation', and 'Promotion'. At the bottom, there are buttons for 'Appliquer', 'OK', and 'Annuler'.

Colonie	Type	N. Modèle de date (Ctrl+Espace)	Length	Precision	Défaut	Commentaire
id_salarie	String	<input checked="" type="checkbox"/>	13	0		
Anciennete_an	Float	<input checked="" type="checkbox"/>		0		
Distance_domicile_Travail	Integer	<input checked="" type="checkbox"/>	2	0		
Service	String	<input checked="" type="checkbox"/>	15	0		
Work_accident	String	<input checked="" type="checkbox"/>	3	0		
Niveau_de_satisfaction	Integer	<input checked="" type="checkbox"/>	3	0		

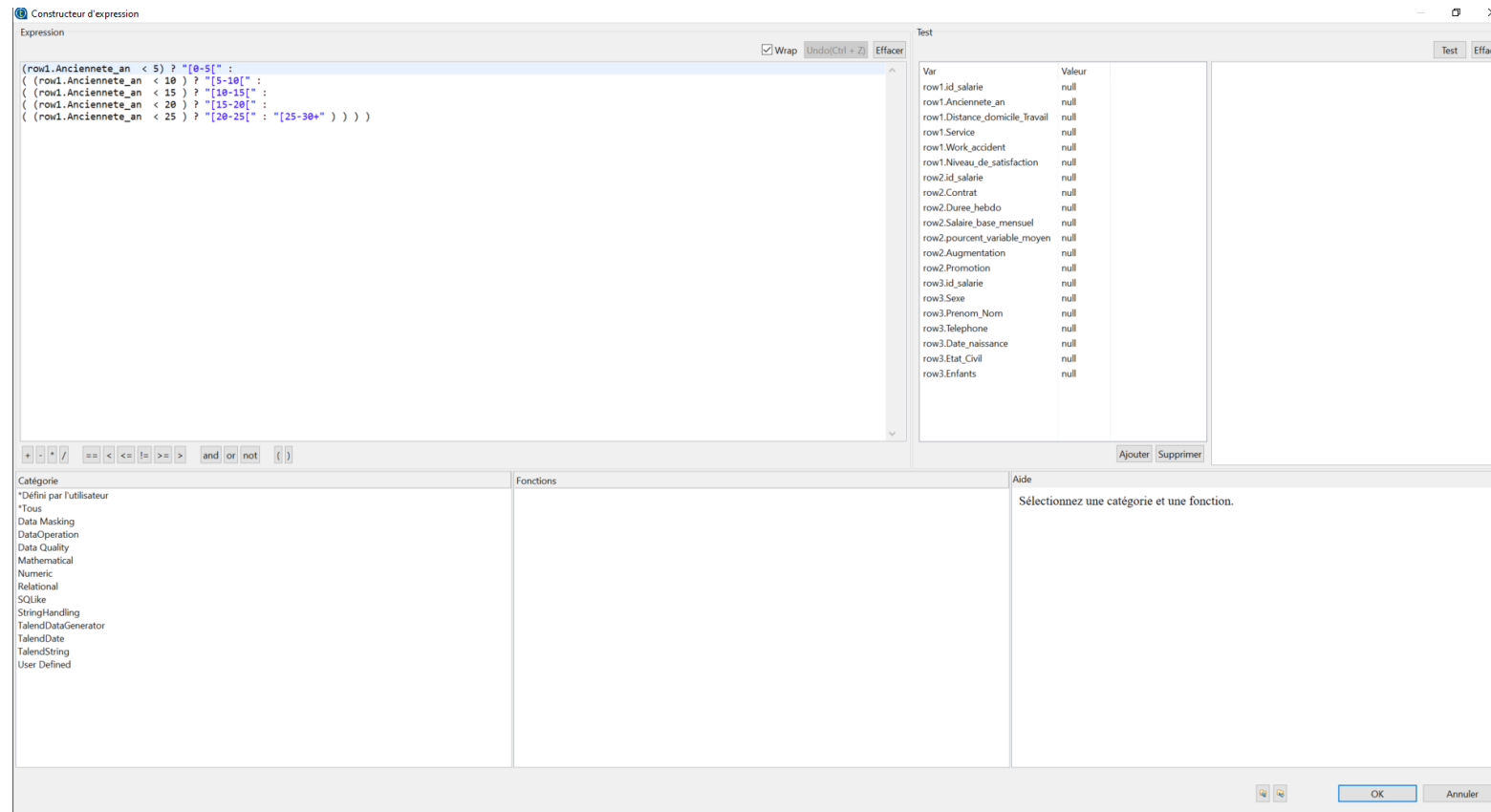
Colonie	Type	N. Modèle de date (Ctrl+Espace)	Length	Precision	Défaut	Commentaire
Service	String	<input checked="" type="checkbox"/>	15	0		
Sexe	Character	<input checked="" type="checkbox"/>	1	0		
Age_binned	String	<input checked="" type="checkbox"/>				
Contrat	String	<input checked="" type="checkbox"/>	3	0		
Duree_hebdo	Integer	<input checked="" type="checkbox"/>	2	0		
Seniority_binned	String	<input checked="" type="checkbox"/>				
Salaire_total	Float	<input checked="" type="checkbox"/>	9			
Augmentation	Integer	<input checked="" type="checkbox"/>	1	0		
Promotion	Integer	<input checked="" type="checkbox"/>	1	0		

- Seules les données utiles aux analyses sont conservées
- Création de 3 nouvelles variables au moyen du constructeur d'expression

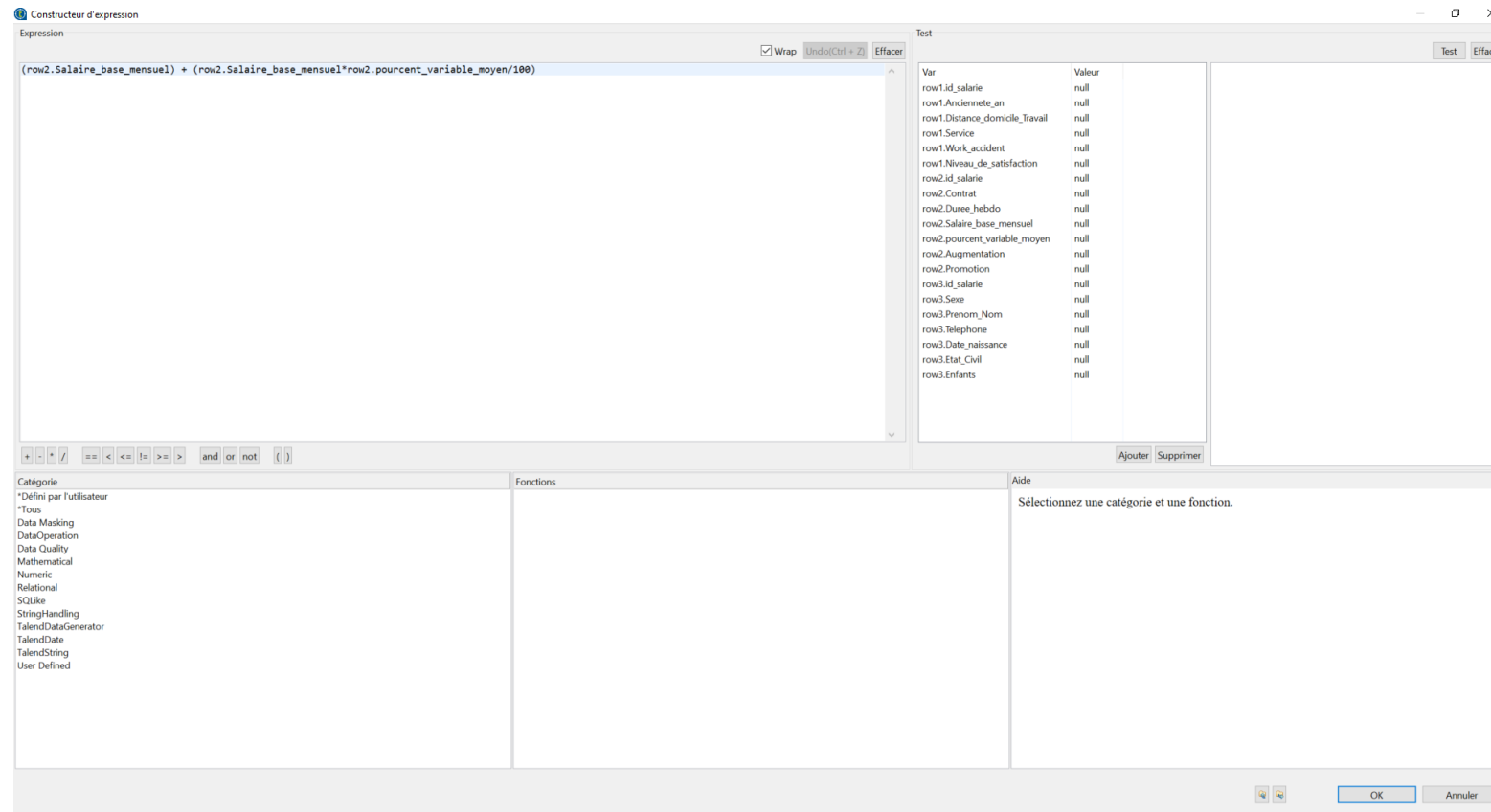
tMap constructeur d'expression : calcul de l'âge et discrétisation



tMap constructeur d'expression : discrétisation de la variable ancienneté

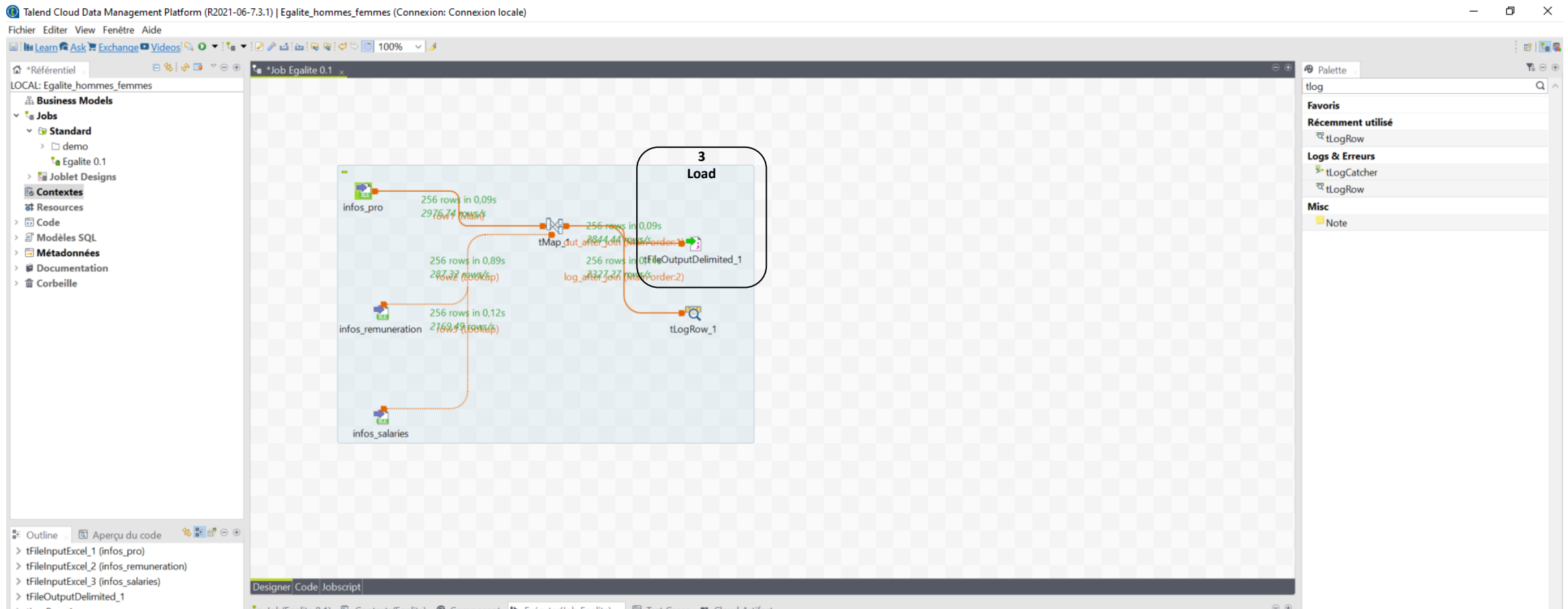


tMap constructeur d'expression : calcul du salaire total



3 Load

3 Load : tFileOutputDelimited



tFileOutputDelimited : création d'un fichier .CSV

The screenshot displays the Talend Cloud Data Management Platform interface. The main workspace shows a job configuration for 'Gender_equality' with a 'tFileOutputDelimited_1' component highlighted. A callout box provides the configuration details for this component.

Choix des paramètres

Paramètres simples	
Type de propriété	Built-In
<input type="checkbox"/> Utiliser le flux de sortie	
Nom de fichier "C:/Users/Celine/workspace/GENDER_EQUALITY/out.csv"	
Séparateur de lignes "\n"	
Séparateur de champs ";"	
<input type="checkbox"/> Ecrire après <input checked="" type="checkbox"/> Inclure l'en-tête <input type="checkbox"/> Compresser en tant que fichier zip	
Règles de validation	
Schéma	Built-In
<input type="checkbox"/> Modifier le schéma <input type="checkbox"/> Sync colonnes	

tFileOutputDelimited : encodage UTF-8

The screenshot displays the Talend Cloud Data Management Platform interface. The main workspace shows a job configuration for 'Gender_equality_indicators 0.1'. The job flow includes several input components (infos_pro, infos_salaries, infos_remuneration) feeding into a tMap_1 component, which then feeds into a tLogRow_1 component, and finally into a tFileOutputDelimited_1 component. A callout box highlights the tFileOutputDelimited_1 component in the job flow and its properties window.

The properties window for tFileOutputDelimited_1 is shown in the bottom right, with the 'Paramètres avancés' (Advanced Parameters) tab selected. The 'Encodage' (Encoding) dropdown is set to 'UTF-8', which is highlighted by a red circle and a red arrow. The text 'Choix de l'encodage' (Choice of encoding) is placed next to the dropdown.

Paramètres avancés

- ☐ Séparateur avancé (pour les nombres)
- ☐ Options CSV
- ☒ Créer le répertoire s'il n'existe pas.
- ☐ Diviser la sortie dans plusieurs fichiers
- ☐ Personnaliser la taille de la mémoire tampon
- ☐ Sortie en mode ligne
- ☐ Ne pas générer de fichier vide
- ☐ Retourner une erreur si le fichier existe déjà
- ☐ Statistiques du tStatCatcher
- ☐ Exécuter parallèlement

Encodage: UTF-8

3 Load :tLog, visualisation de la table de sortie dans l'interface Talend ®

