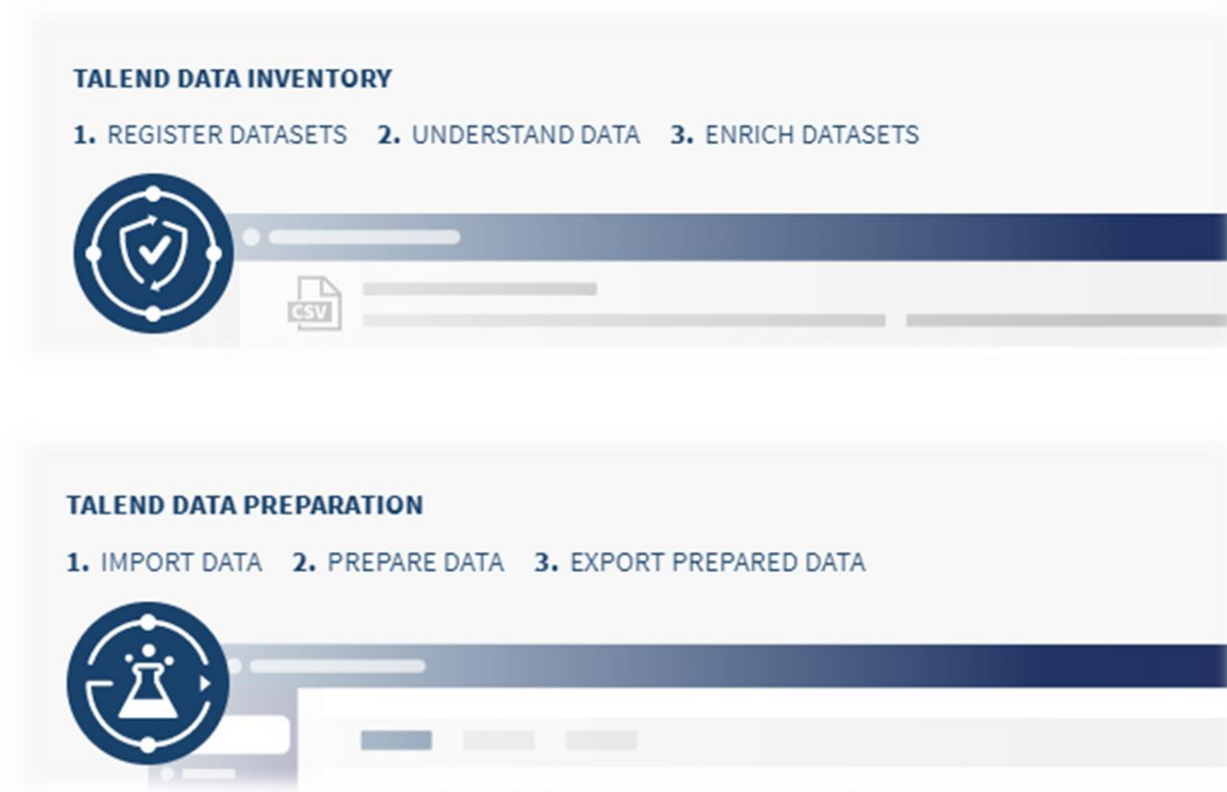


Utilisation de Talend Cloud® pour la création d'un fichier .CSV avec données anonymisées



Objectif

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

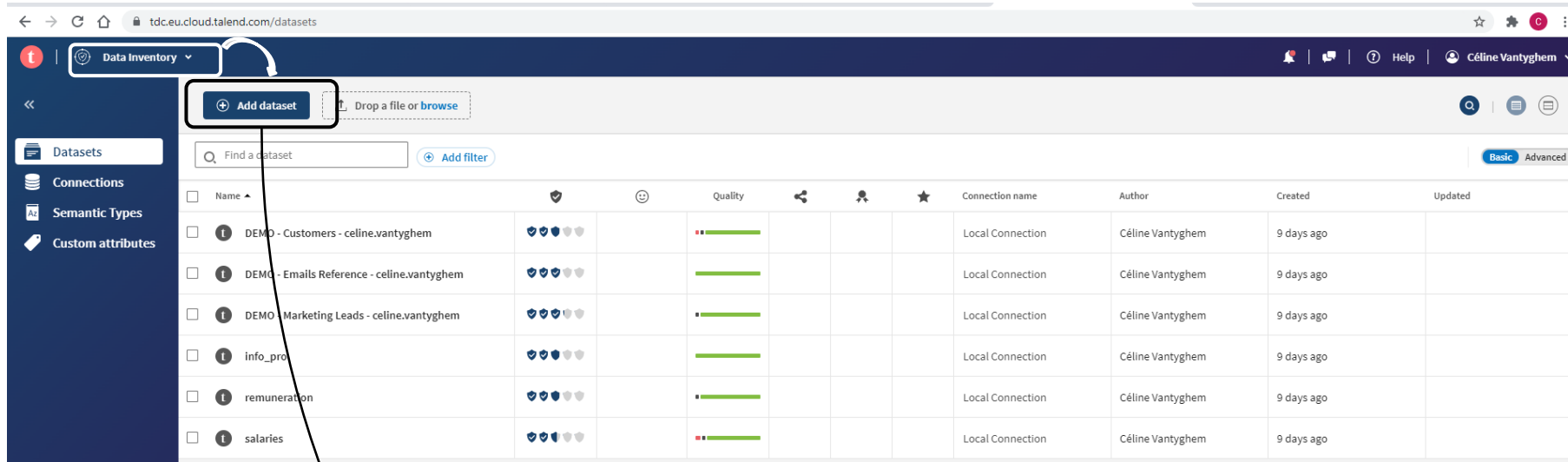
Objectif : créer un fichier CSV avec les données issues des trois fichiers source, dans le respect du RGPD. Ce fichier sera utilisé pour la réalisation d'un diagnostic égalité professionnelle hommes-femmes (embauche, qualification, promotion, conditions de travail , rémunération...)

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

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

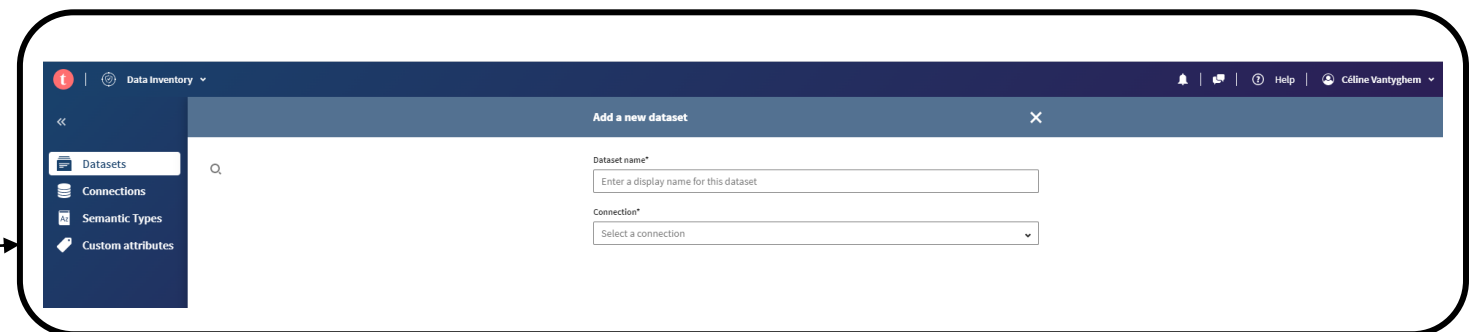
=> création de nouvelles variables (salaire total...)

Ajouter les datasets en utilisant l'application Data Inventory



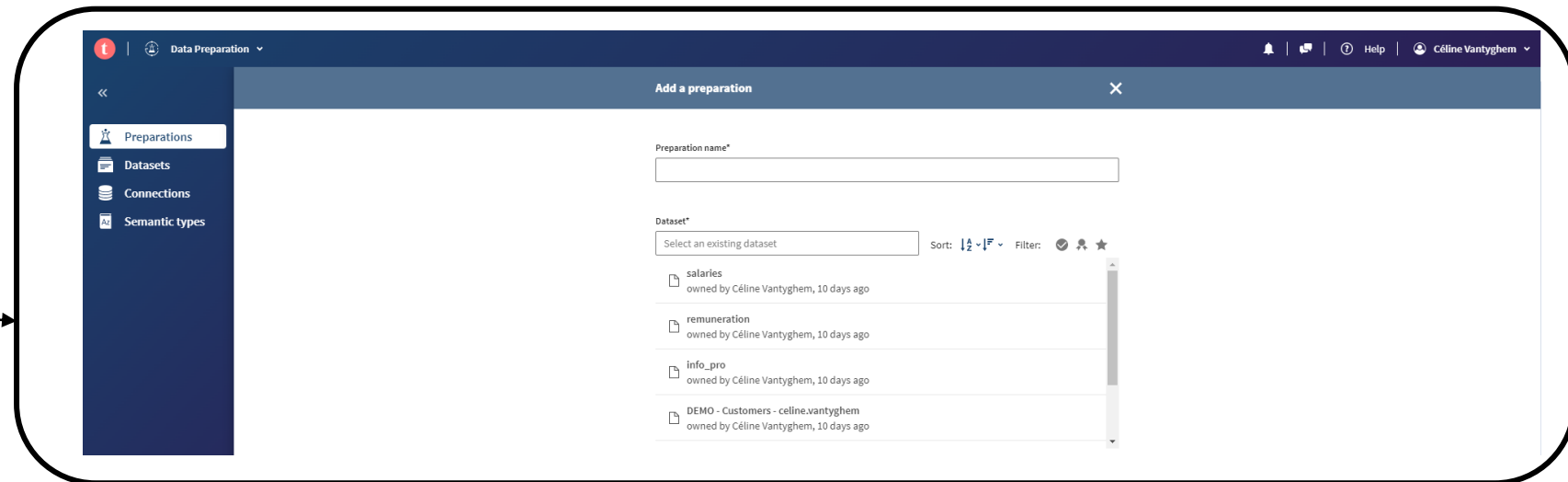
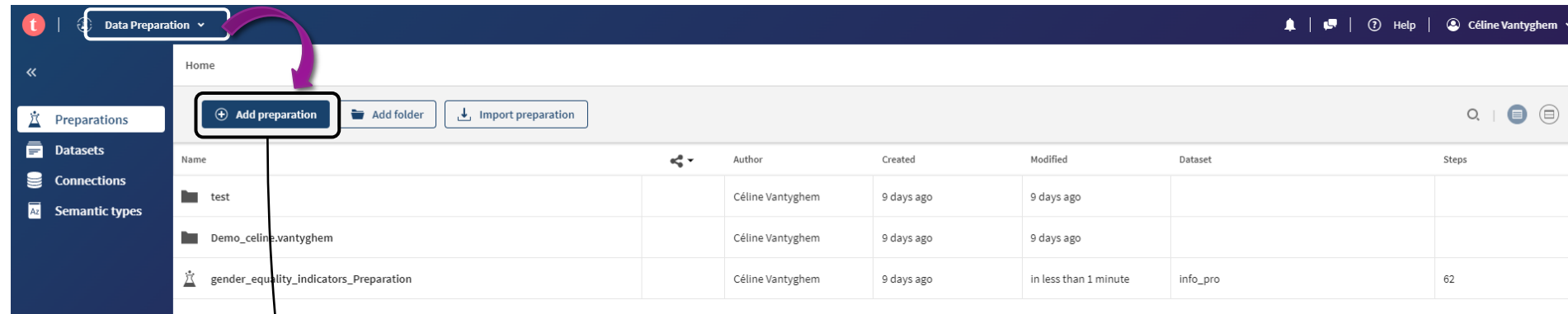
The screenshot shows the Talend Data Inventory web application interface. The left sidebar contains navigation links: Datasets, Connections, Semantic Types, and Custom attributes. The main area displays a table of datasets. A red box highlights the 'Add dataset' button in the top left corner of the main area. A curved arrow points from this button to the 'Add a new dataset' modal window shown in the bottom right.

| Name | Quality | Connection name | Author | Created | Updated |
|---|---------|------------------|-----------------|------------|---------|
| DEMO - Customers - celine.vantyghe | ■■■■■ | Local Connection | Céline Vantyghe | 9 days ago | |
| DEMO - Emails Reference - celine.vantyghe | ■■■■■ | Local Connection | Céline Vantyghe | 9 days ago | |
| DEMO - Marketing Leads - celine.vantyghe | ■■■■■ | Local Connection | Céline Vantyghe | 9 days ago | |
| info_pro | ■■■■■ | Local Connection | Céline Vantyghe | 9 days ago | |
| remuneration | ■■■■■ | Local Connection | Céline Vantyghe | 9 days ago | |
| salaries | ■■■■■ | Local Connection | Céline Vantyghe | 9 days ago | |

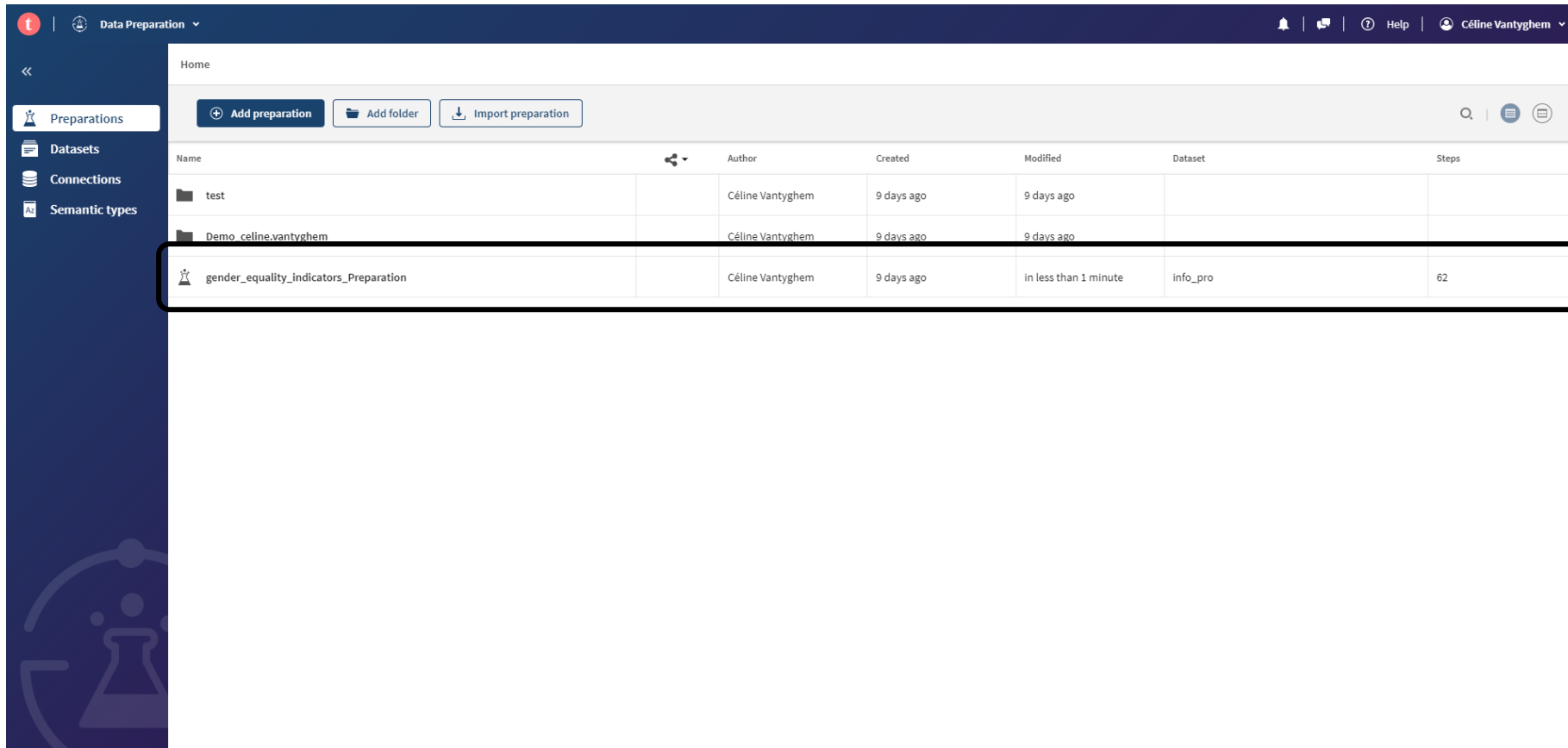


The modal window titled 'Add a new dataset' is open. It contains a search bar on the left and two input fields on the right: 'Dataset name*' with a placeholder 'Enter a display name for this dataset' and 'Connection*' with a dropdown menu labeled 'Select a connection'.

Créer une nouvelle préparation en utilisant l'application Data Preparation



Cliquer sur la préparation créée dans l'application Data Preparation



The screenshot shows the Talend Data Preparation application interface. The left sidebar contains navigation links: Preparations, Datasets, Connections, and Semantic types. The main area displays a table of preparations. The table has columns for Name, Author, Created, Modified, Dataset, and Steps. The preparation 'gender_equality_indicators_Preparation' is highlighted with a red box.

| Name | Author | Created | Modified | Dataset | Steps |
|--|-----------------|------------|-----------------------|----------|-------|
| test | Céline Vantghem | 9 days ago | 9 days ago | | |
| Demo celine.vantghem | Céline Vantghem | 9 days ago | 9 days ago | | |
| gender_equality_indicators_Preparation | Céline Vantghem | 9 days ago | in less than 1 minute | info_pro | 62 |

Interface Data Preparation

3 Flux des opérations effectuées :

les étapes s'ajoutent une à une. L'ordre des opérations peut être modifié par glissé-déposé. Chaque opération peut être annulée. Ce workflow pourra s'appliquer à de nouveaux datasets

- 1 **Lookup** done with dataset remuneration. Join has been set between id_salarie and id_salarie. The columns Contrat, Durée hebdo and 4 other(s) have been added.
- 2 **Lookup** done with dataset salaries. Join has been set between id_salarie and id_salarie. The columns Sexe and Date_naissance have been added.
- 3 **Fill invalid cells with value** on column Sexe
- 4 **Magic fill** on column Date_naissance
- 5 **Rename column** on column Date_naissance_magic_fill
- 6 **Generate sequence** on column Annee_naissance
- 7 **Add, multiply, subtract or divide** on column Annee_naissance_sequence
- 8 **Rename column** on column Annee_naissance_sequence
- 9 **Change data type** on column Contrat
- 10 **Magic fill** on column Salaire base mensuel
- 11 **Delete column** on column Salaire base mensuel

1 Sélectionner une ou plusieurs colonnes ou une ou plusieurs lignes

| Filters | | | | | | | | | | |
|--------------|-----------------|------|------------|---------|-------------|------------------|---------------|----|-----|--|
| Add a filter | | | | | | | | | | |
| | Service | Sexe | Age_binned | Contrat | Durée hebdo | Seniority_binned | Salaire total | Au | int | |
| 1 | Marketing | F | 60-+ | CDI | | 35 | 7439.7595 | | | |
| 2 | Commercial | F | 60-+ | CDI | | 24 | 6566.1024 | | | |
| 3 | RH | F | 41-50 | CDI | | 32 | 10544.332 | | | |
| 4 | Compta Finances | F | 30-40 | CDI | | 35 | 9493.9498 | | | |
| 5 | Consultant | F | 41-50 | CDI | | 35 | 8545.9559 | | | |
| 6 | RH | F | 51-60 | CDI | | 35 | 9241.8943 | | | |
| 7 | Commercial | M | 41-50 | CDI | | 35 | 2068.274 | | | |
| 8 | Consultant | F | 41-50 | CDI | | 35 | 13499.85 | | | |
| 9 | Commercial | F | 51-60 | CDI | | 35 | 11051.768 | | | |
| 10 | Commercial | F | 41-50 | CDI | | 35 | 5653.1967 | | | |
| 11 | Consultant | F | 51-60 | CDI | | 35 | 11688.3392 | | | |
| 12 | RH | F | 60-+ | CDI | | 24 | 4446.7488 | | | |
| 13 | Commercial | F | 51-60 | CDI | | 35 | 7149.9268 | | | |
| 14 | RH | F | 60-+ | CDI | | 35 | 7357.5063 | | | |
| 15 | Compta Finances | M | 60-+ | CDI | | 35 | 3437.4384 | | | |
| 16 | Marketing | M | 41-50 | CDI | | 35 | 9842.6965 | | | |
| 17 | Consultant | M | 51-60 | CDI | | 28 | 3038.8176 | | | |

2 Appliquer une opération parmi la liste des opérations disponibles
Ex : search and replace

Service

Column Row Table

Q Filter

SUGGESTIONS

- Change to lower case ...
- Change to title case ...
- Change to upper case ...
- Magic fill ...
- Search and replace ...

BOOLEAN

- Negate value ...

Chart Value Pattern Advanced

Row count *

0 20 40 60 80

Consultant

Commercial

RH

Compta Finances

Marketing

R&D

Jointure

La jointure avec d'autres datasets s'effectue avec ce bouton après avoir sélectionné la colonne qui sera utilisée comme clé

The screenshot displays the Talend Data Preparation interface. The main window shows a table with 17 rows and 10 columns. The columns are: Service (text), Sexe (Gender), Age_binned (text), Contrat (text), Durée hebdo (integer), Seniority_binned (text), Salaire total (decimal), and Aug (integer). The table is titled "gender_equality_indicators_Preparation" and the dataset is "info_pro".

On the left, a sidebar lists 11 transformation steps:

- 1 Lookup done with dataset remuneration. Join has been set between id_salarie and id_salarie. The columns Contrat, Durée hebdo and 4 other(s) have been added.
- 2 Lookup done with dataset salaries. Join has been set between id_salarie and id_salarie. The columns Sexe and Date_naissance have been added.
- 3 Fill invalid cells with value on column Sexe
- 4 Magic fill on column Date_naissance
- 5 Rename column on column Date_naissance_magic_fill
- 6 Generate sequence on column Annee_naissance
- 7 Add, multiply, subtract or divide on column Annee_naissance_sequence
- 8 Rename column on column Annee_naissance_sequence
- 9 Change data type on column Contrat
- 10 Magic fill on column Salaire base mensuel
- 11 Delete column on column Salaire base mensuel

On the right, a "Service" panel shows a "Filter" dropdown and a "SUGGESTIONS" list. Below this, a "Chart" panel displays a horizontal bar chart titled "Row count" showing the distribution of the "Service" column. The chart has a y-axis with categories: Consultant, Commercial, RH, Compta Finances, Marketing, and R&D. The x-axis represents the row count, ranging from 0 to 80.

A red circle highlights a button in the top right corner of the interface, which is used for joining datasets. A red arrow points from the text box above to this button.

Statistiques

gender_equality_indicators_Preparation
Dataset: info_pro

256/256

Filters

Add a filter... [Add filter](#)

| | Service text | Sexe Gender | Age_binned text | Contrat text | Durée hebdo integer | Seniority_binned text | Salaire total decimal | Aug int |
|----|-----------------|----------------|--------------------|-----------------|------------------------|--------------------------|--------------------------|------------|
| 1 | Marketing | F | 60-+ | CDI | | 35 | [10-15[| 7439.7505 |
| 2 | Commercial | F | 60-+ | CDI | | 24 | [10-15[| 6566.1024 |
| 3 | RH | F | 41-50 | CDI | | 32 | [10-15[| 10544.352 |
| 4 | Compta Finances | F | 30-40 | CDI | | 35 | [10-15[| 9493.9428 |
| 5 | Consultant | F | 41-50 | CDI | | 35 | [5-10[| 8545.9959 |
| 6 | RH | F | 51-60 | CDI | | 35 | [5-10[| 9241.8953 |
| 7 | Commercial | M | 41-50 | CDI | | 35 | [10-15[| 2068.274 |
| 8 | Consultant | F | 41-50 | CDI | | 35 | [5-10[| 13499.85 |
| 9 | Commercial | F | 51-60 | CDI | | 35 | [15-20[| 11051.768 |
| 10 | Commercial | F | 41-50 | CDI | | 35 | [20-25[| 5653.1967 |
| 11 | Consultant | F | 51-60 | CDI | | 35 | [20-25[| 11688.3392 |
| 12 | RH | F | 60-+ | CDI | | 24 | [5-10[| 4446.7488 |
| 13 | Commercial | F | 51-60 | CDI | | 35 | [20-25[| 7149.9268 |
| 14 | RH | F | 60-+ | CDI | | 35 | [15-20[| 7357.5063 |
| 15 | Compta Finances | M | 60-+ | CDI | | 35 | [20-25[| 3437.4384 |
| 16 | Marketing | M | 41-50 | CDI | | 35 | [20-25[| 9842.6965 |
| 17 | Consultant | M | 51-60 | CDI | | 28 | [10-15[| 3038.8176 |

Service

Column Row Table

Filter

SUGGESTIONS

Change to lower case ...

Change to title case ...

Change to upper case ...

Magic fill ...

Search and replace ...

BOOLEAN

Negate value ...

Chart Value Pattern Advanced

Row count *

0 20 40 60 80

Consultant

Commercial

RH

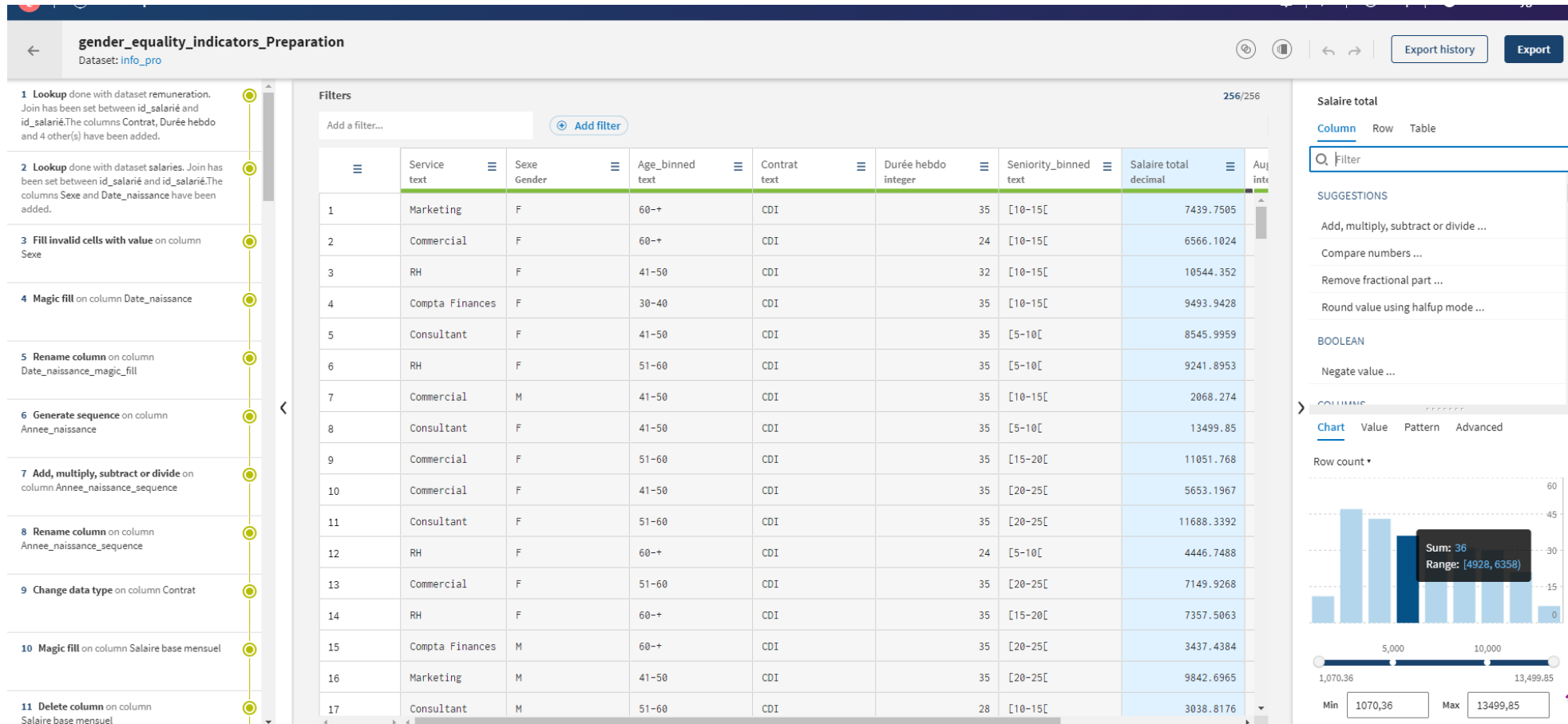
Compta Finances

Marketing

R&D

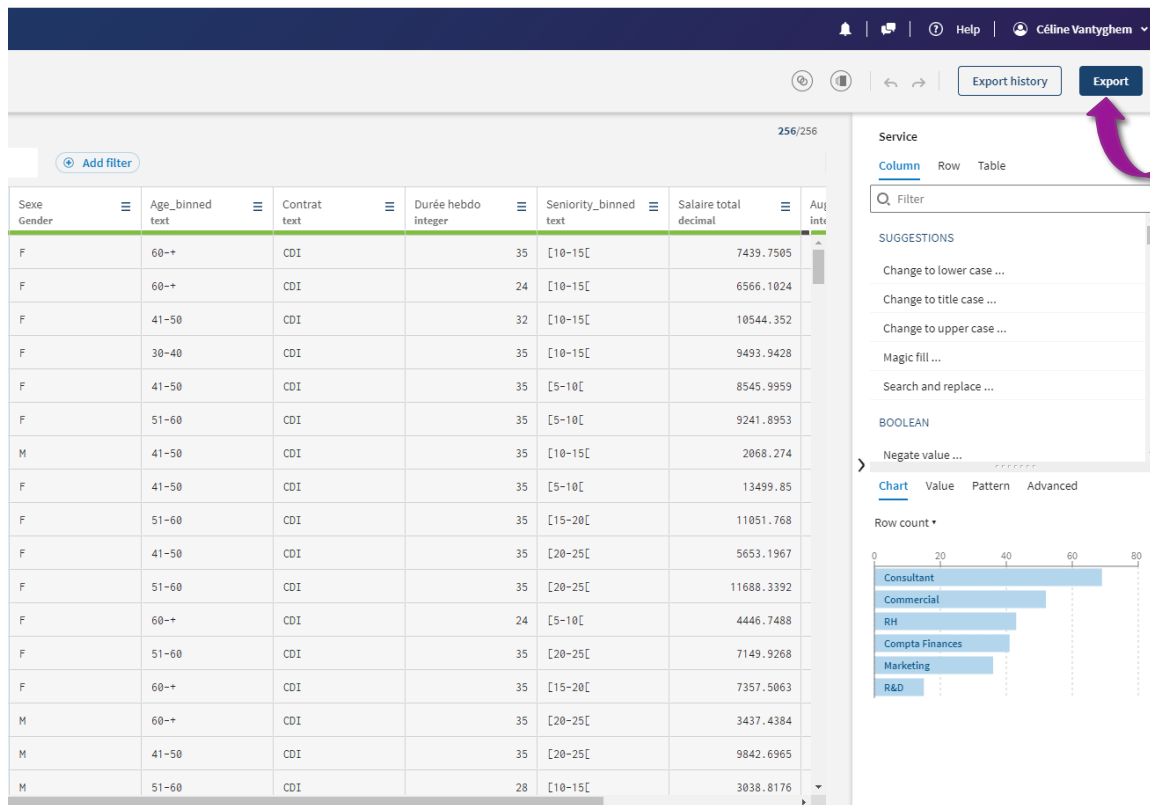
Des statistiques sont calculées automatiquement pour chaque variable
Exemple pour une variable catégorielle

Statistiques



Des statistiques sont calculées automatiquement pour chaque variable.
Exemple pour une variable numérique

Export du fichier



The screenshot shows the Talend Data Studio interface. At the top, there's a header bar with a user profile 'Céline Vantyghe' and a navigation bar. Below the header, there's a toolbar with buttons for 'Export history' and 'Export'. The main area displays a data table with columns: Sexe, Age_binned, Contrat, Durée hebdo, Seniority_binned, Salaire total, and Aug. The table contains 256 rows of data. A sidebar on the right shows a 'Service' panel with tabs for 'Column', 'Row', and 'Table'. The 'Column' tab is active, showing a list of columns and a 'Filter' section. Below the filter, there are 'SUGGESTIONS' for text manipulation (e.g., 'Change to lower case ...'), 'BOOLEAN' operations (e.g., 'Negate value ...'), and a 'Chart' section with a bar chart showing row counts for different categories like 'Consultant', 'Commercial', 'RH', etc.

| Sexe | Age_binned | Contrat | Durée hebdo | Seniority_binned | Salaire total | Aug |
|------|------------|---------|-------------|------------------|---------------|-----|
| F | 60++ | CDI | | 35 [10-15] | 7439.7505 | |
| F | 60++ | CDI | | 24 [10-15] | 6566.1024 | |
| F | 41-50 | CDI | | 32 [10-15] | 10544.352 | |
| F | 30-40 | CDI | | 35 [10-15] | 9493.9428 | |
| F | 41-50 | CDI | | 35 [5-10] | 8545.9959 | |
| F | 51-60 | CDI | | 35 [5-10] | 9241.8953 | |
| M | 41-50 | CDI | | 35 [10-15] | 2068.274 | |
| F | 41-50 | CDI | | 35 [5-10] | 13499.85 | |
| F | 51-60 | CDI | | 35 [15-20] | 11051.768 | |
| F | 41-50 | CDI | | 35 [20-25] | 5653.1967 | |
| F | 51-60 | CDI | | 35 [20-25] | 11688.3392 | |
| F | 60++ | CDI | | 24 [5-10] | 4446.7488 | |
| F | 51-60 | CDI | | 35 [20-25] | 7149.9268 | |
| F | 60++ | CDI | | 35 [15-20] | 7357.5063 | |
| M | 60++ | CDI | | 35 [20-25] | 3437.4384 | |
| M | 41-50 | CDI | | 35 [20-25] | 9842.6965 | |
| M | 51-60 | CDI | | 28 [10-15] | 3038.8176 | |

L'export de la table finale est réalisé en cliquant sur le bouton Export

Les paramètres de l'export sont choisis via la fenêtre de configuration

Export

☐ Current sample ☒ All data

- ☐ Local CSV file
- ☐ Local XLSX file
- ☐ Local Tableau file
- ☐ Amazon S3

Cancel

Confirm

Export du fichier

The screenshot shows the Talend Data Preparation interface. At the top, there's a header bar with a user profile 'Céline Vantigham' and a 'Help' button. Below the header, there's a toolbar with buttons for 'Export history' and 'Export'. The main area displays a data table with columns: Sexe/Gender, Age_binned/text, Contrat/text, Durée hebdo/integer, Seniority_binned/text, Salaire total/decimal, and Aug/int. The table contains 256 rows. To the right of the table, there's a sidebar with a 'Service' tab and a 'Filter' search bar. Below the search bar, there are several sections: 'SUGGESTIONS' (Change to lower case, Change to title case, Change to upper case, Magic fill, Search and replace), 'BOOLEAN' (Negate value), and 'Chart' (Value, Pattern, Advanced). The 'Chart' section is currently selected, showing a horizontal bar chart with categories: Consultant, Commercial, RH, Compta Finances, Marketing, and R&D.

| Sexe Gender | Age_binned text | Contrat text | Durée hebdo integer | Seniority_binned text | Salaire total decimal | Aug int |
|----------------|--------------------|-----------------|------------------------|--------------------------|--------------------------|------------|
| F | 60++ | CDI | 35 | [10-15[| 7439.7505 | |
| F | 60++ | CDI | 24 | [10-15[| 6566.1024 | |
| F | 41-50 | CDI | 32 | [10-15[| 10544.352 | |
| F | 30-40 | CDI | 35 | [10-15[| 9493.9428 | |
| F | 41-50 | CDI | 35 | [5-10[| 8545.9959 | |
| F | 51-60 | CDI | 35 | [5-10[| 9241.8953 | |
| M | 41-50 | CDI | 35 | [10-15[| 2068.274 | |
| F | 41-50 | CDI | 35 | [5-10[| 13499.85 | |
| F | 51-60 | CDI | 35 | [15-20[| 11051.768 | |
| F | 41-50 | CDI | 35 | [20-25[| 5653.1967 | |
| F | 51-60 | CDI | 35 | [20-25[| 11688.3392 | |
| F | 60++ | CDI | 24 | [5-10[| 4446.7488 | |
| F | 51-60 | CDI | 35 | [20-25[| 7149.9268 | |
| F | 60++ | CDI | 35 | [15-20[| 7357.5063 | |
| M | 60++ | CDI | 35 | [20-25[| 3437.4384 | |
| M | 41-50 | CDI | 35 | [20-25[| 9842.6965 | |
| M | 51-60 | CDI | 28 | [10-15[| 3038.8176 | |

L'historique des fichiers créés est disponible en cliquant sur le bouton Export history

The screenshot shows the 'Export history' window in the Talend Data Preparation interface. The window title is 'Export history' and the subtitle is 'Preparation: gender_equality_indicators_Preparation'. It displays a list of successful exports. Each entry includes a status (Successful), the creator (Céline VANTIGHM), the start time, the time spent, and the file format (CSV). The first entry shows a file named 'gender_equality_indicators_Preparation.csv' with a time spent of 00:00.

| Status | Created by | Start | Time spent | Format |
|------------|-----------------|------------------|------------|--------|
| Successful | Céline VANTIGHM | 20-10-2021 17:05 | 00:00:42 | CSV |
| Successful | Céline VANTIGHM | 19-10-2021 14:44 | 00:00:45 | CSV |
| Successful | Céline VANTIGHM | 19-10-2021 12:54 | 00:00:37 | CSV |