talend

Connecting the Data-Driven Enterprise

Getting Started with Talend Data Stewardship

6.4.1

Contents

Copyright

Adapted for 6.4.1. Supersedes previous releases.

Publication date: June 29th, 2017

Merging data with Talend Data Stewardship

Talend Data Stewardship is a tool you can use to manage data assets. It organizes the interactions on data whenever human intervention is required to collaborate to data curation, arbitration or validation.

The core concepts of Talend Data Stewardship are campaigns and tasks. It comes with two predefined roles namely: campaign owners and data stewards.

A campaign is the main unit of work for campaign owners. It contains all the required configuration assets that are determined by the campaign owner:

- what are the tasks about (data structure, validation constraints, etc.)?
- what do data stewards have to do to resolve the campaign tasks (task type)?
- which data stewards work on the campaign tasks (campaign participants)?
- how data stewards collaborate to resolve the campaign tasks (campaign workflow)?

A task is the main unit of work for data stewards. A task belongs to a campaign and is assigned to a data steward. It has a lifecycle where it passes through different states according to the workflow defined in the campaign.

One of the solutions Talend Data Stewardship provides is to match, cleanse and master data using a **Merging** campaign.

The use case described here uses a **Merging** campaign to illustrate the aspects of reconciling data coming from different sources. However, Talend Data Stewardship supports additional campaign types.

For further information, see the online publication about Talend Data Stewardship use cases on Talend Help Center (https://help.talend.com).

In a **Merging** campaign, you must consider two aspects:

- How do you identify the match groups which group potentially duplicate records together? This question is resolved through using a Talend Job in the Studio.
- How do you pick the best attribute values from the data sources and presents the most accurate and reliable master records for consumptions by users and systems? This issue is resolved through the **Merging** campaign in the web application.

This use case helps you getting started with Talend Data Stewardship. To replicate the example and use the exact client data, we assume that:

- An administrator has installed and launched Talend Data Stewardship. For more information, see the Talend Administration Center Installation Guide.
- An administrator has created Talend Data Stewardship users and assigned them roles in Talend Administration Center. For further information, see the online publication about creating data stewardship users on Talend Help Center (https://help.talend.com).
- A campaign owner has downloaded the input data and the Talend Job used in this example. They can be used to load tasks in the **Merging** campaign once it is created.

  Retrieve the `tds_gettingstarted_source_files.zip` file from the **Downloads** tab of the online version of this page at https://help.talend.com.

Adding a Merging campaign to reconciliate data

As a campaign owner, you can create a campaign to determine the structure of the data to be managed, the actions to be taken on data and what data stewards to work on what tasks.

You can set permission per attribute and per data steward to define who can view or edit which attributes. You can also decide how campaign participants will collaborate to resolve the campaign tasks. A campaign can be owned by one or several campaign owners.

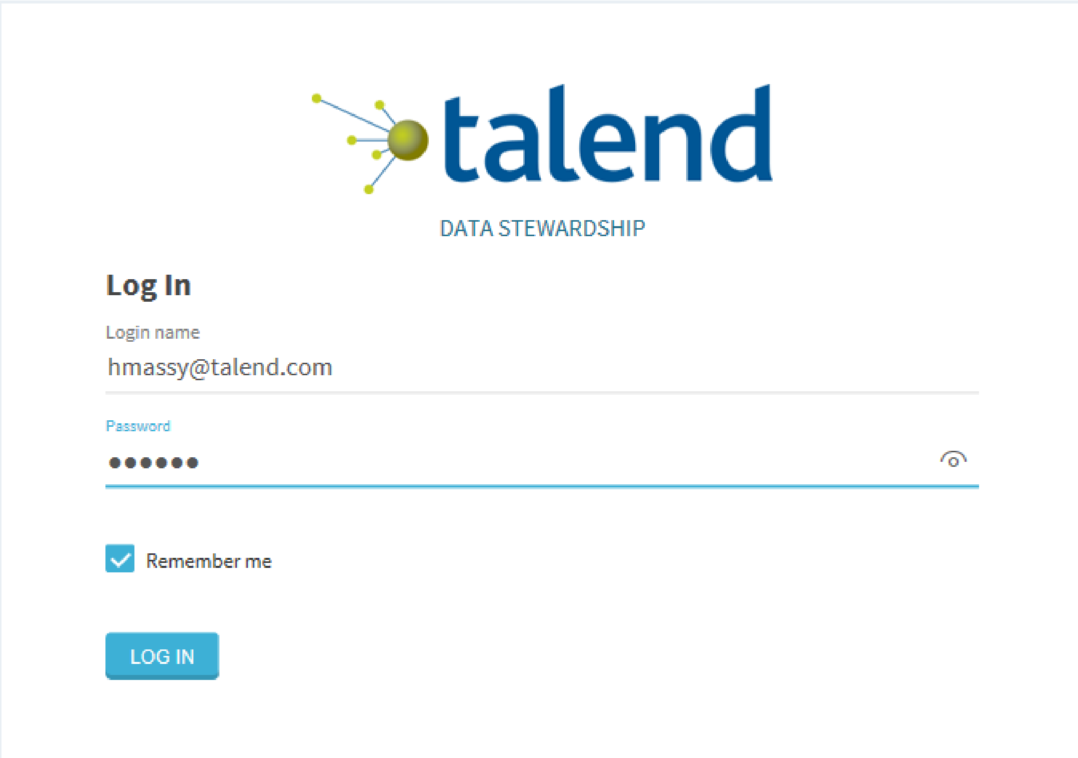
Logging in to Talend Data Stewardship for the first time

Once Talend Data Stewardship is installed and launched, you can use your login information to get started.

A certain sequence need to be followed in order to successfully start Talend Data Stewardship and its dependencies.

For further information, see the online publication on start and stop sequences on Talend Help Center (https://help.talend.com).

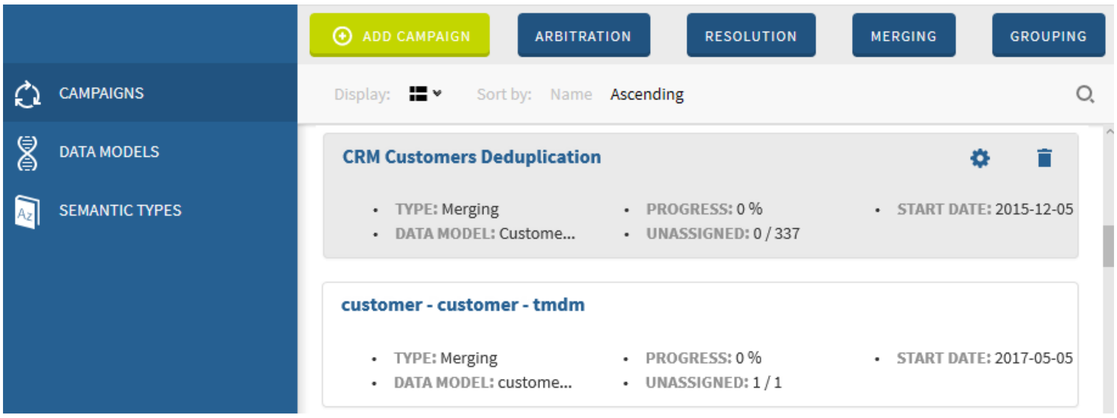
1. Open a web browser and navigate to the URL `http://<hostname>:<port>` after replacing the host and port names with the appropriate values depending on your installation.
   For example, use `http://localhost:19999` if the default parameters have been used for installing Talend Data Stewardship. Otherwise, contact the administrator who installed the application to get the host and port names.
2. Enter the credentials you received from the Administrator.



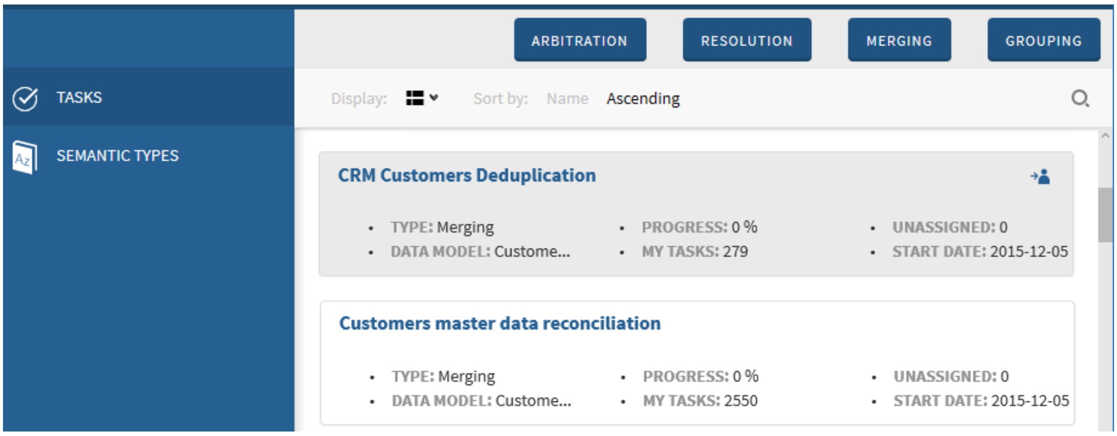
3. Click **Log in**.

   The home page is opened. The options available in this page depends on your role.

   Home page of campaign owner:



Home page of data stewards:



Defining a data model for the campaign

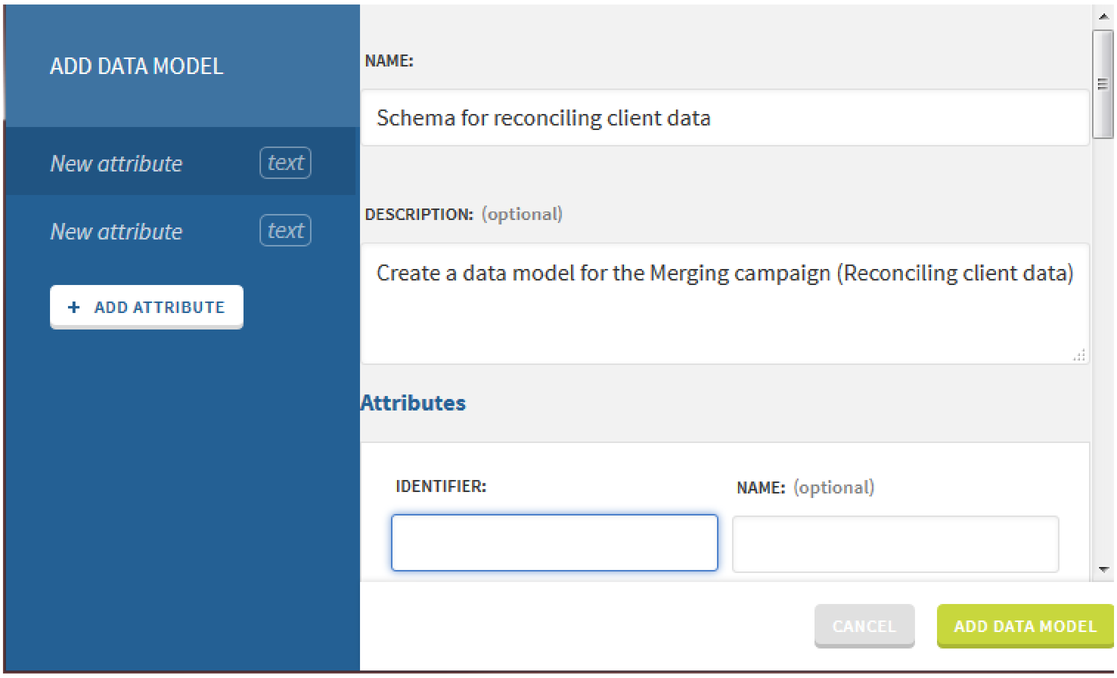
In this example, you need to create a data model to determine the structure of the data to be managed in a campaign. You will use this data model in a `Reconciling client data` campaign you create to allow data stewards to merge redundant client data coming from different systems.

Talend Data Stewardship has data model awareness which makes syntactic and semantic validation of data possible. Therefore, a campaign relies on a data model to ensure that the data matches the expected structure and format.

By defining the structure of data for this **Merging** campaign, you define the attributes in the data model and select their types out of a predefined standard or semantic types.

- An administrator has created Talend Data Stewardship users and assigned them roles in Talend Administration Center. For further information, see the online publication about creating data stewardship users on Talend Help Center (https://help.talend.com).
- You have been assigned a campaign owner role in Talend Administration Center.
- You have accessed Talend Data Stewardship as a campaign owner.

1. In the home page, click **DATA MODELS** > **ADD DATA MODEL**.
2. Enter a name and a description for the new model in the **Name** and **Description** fields respectively.



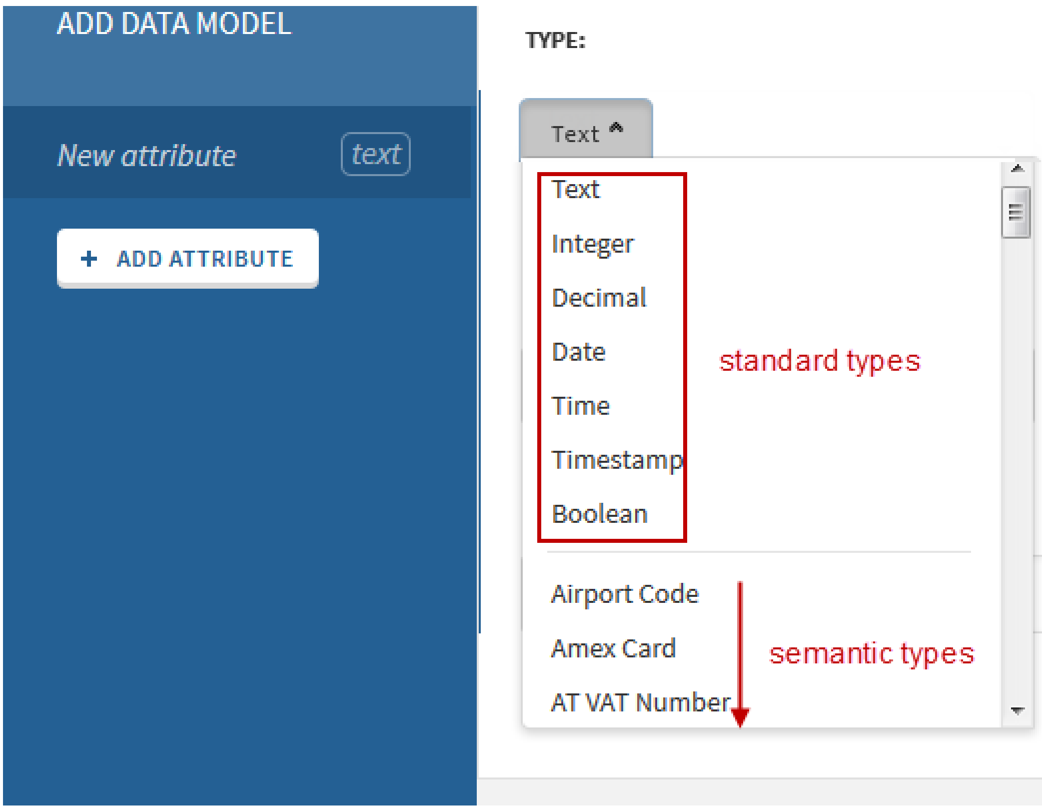
Optional fields are marked as **optional** next to their names

3. In the **Attributes** section, define the columns you want to have in the data model as the following:

a) In the **IDENTIFIER** field, enter the technical identifier for the first column.

b) Enter a name and a description for the column in the corresponding fields, if needed.
What you set in the **NAME** field is the name displayed in the task list. If no name is set, the technical identifier will be displayed.

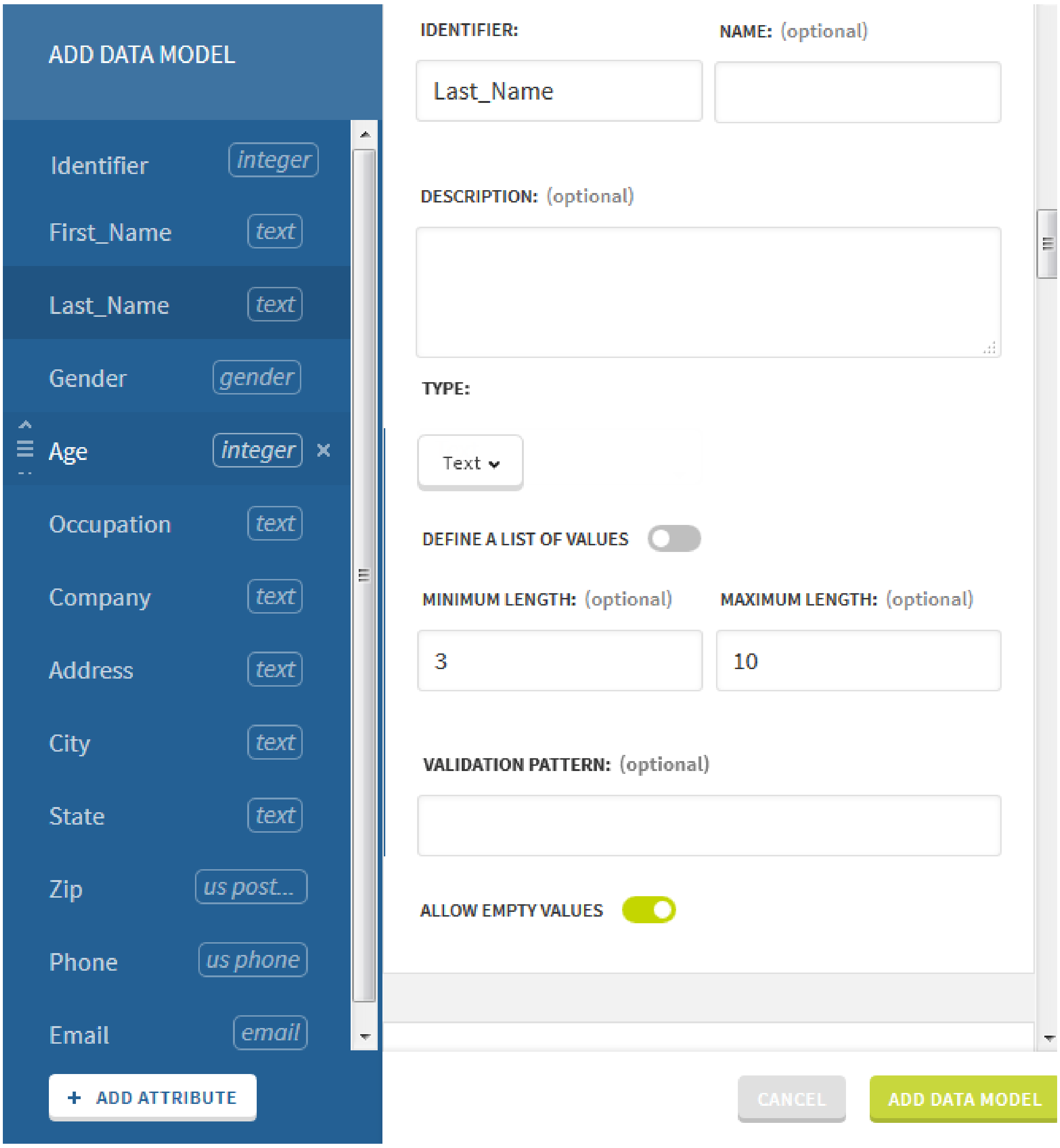
c) From the attribute type list, select the type of the column.



Standard and semantic types are integrated in Talend Data Stewardship by default

- For standard types, additional fields are displayed or hidden according to the type you select. These fields are optional and they enable you to define some constraints on the attribute you define such as defining a minimum and/or maximum length or defining a pattern against which to validate the attribute.
- For semantic types, you can use the Talend Dictionary Service to manage the semantic types. However, the availability of this service depends on the license you have.

4. Click the switch next to **DEFINE A LIST OF VALUES** to display fields where you can set specific values for the attribute.
   Any values that are not in this list are marked as invalid in the task list.
5. Click the switch next to **ALLOW EMPTY VALUES** to disable the load of empty fields to Talend Data Stewardship, if needed. This option is enabled by default.
6. Click **ADD ATTRIBUTE** in the left panel and repeat the above steps to create all the columns you need in the data model.

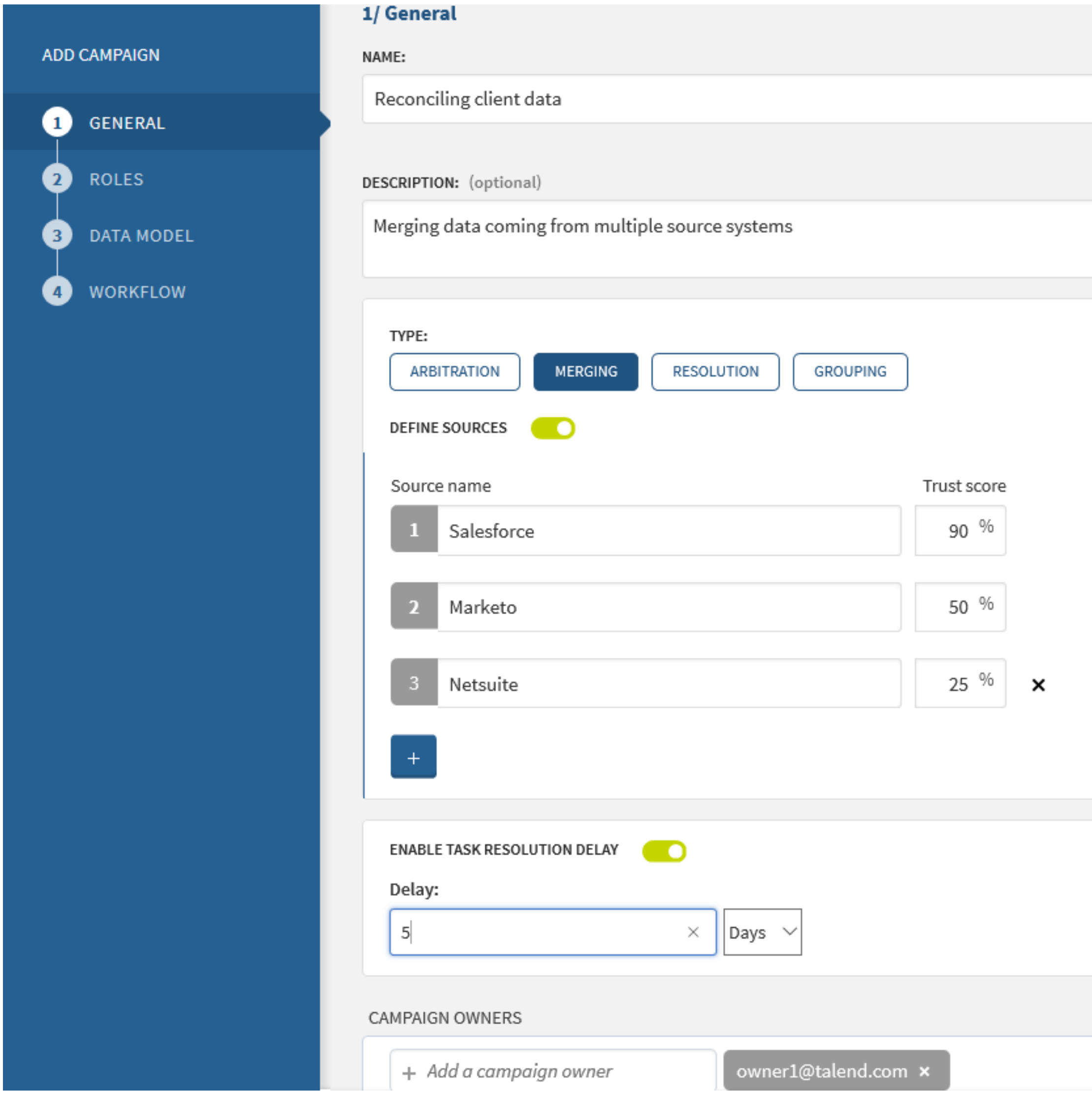


Defining the campaign

The **Merging** campaign used in this example enables you to compare the redundant contact details retrieved from different systems and decide what attribute values define the master records.

Start by defining the campaign metadata.

1. Click **CAMPAIGNS** > **ADD CAMPAIGN**.
2. Enter a name and a description for the new campaign.
3. Select **MERGING** as the campaign type.



4. Click the switch next to **Define Sources** to display a field where you can set the first source name.
5. Use the plus button to add two other fields under sources and set the sources names. Make sure that the sources names do not contain dots and that they do not start with dollar signs.
   Client data here comes from three different sources: Salesforce, Marketo and Netsuite.
6. In each of the **Trust Score** fields, enter the percentage which defines the weight of the value coming from the data source.
   These scores are used by Talend Data Stewardship to merge redundant data and create by-default master records.

   When data stewards access the task list, master records are already defined according to these scores. However, they can manually set survivorship rules per attributes in the data records or enter completely new values when resolving the task.

   In this example, the contact information which comes from Salesforce is initially used by the application to form master records as you gave Salesforce the highest score, while Marketo comes second in the list.

7. Click the switch next to **ENABLE TASK RESOLUTION DELAY** and set the delay to be used to calculate the due dates.

   The due date is calculated by adding the delay you set to the date and time of loading the tasks in the campaign. If you leave this option disabled, the **DUE DATE** column in the task list stays empty and tasks in the campaign do not have resolution delay.

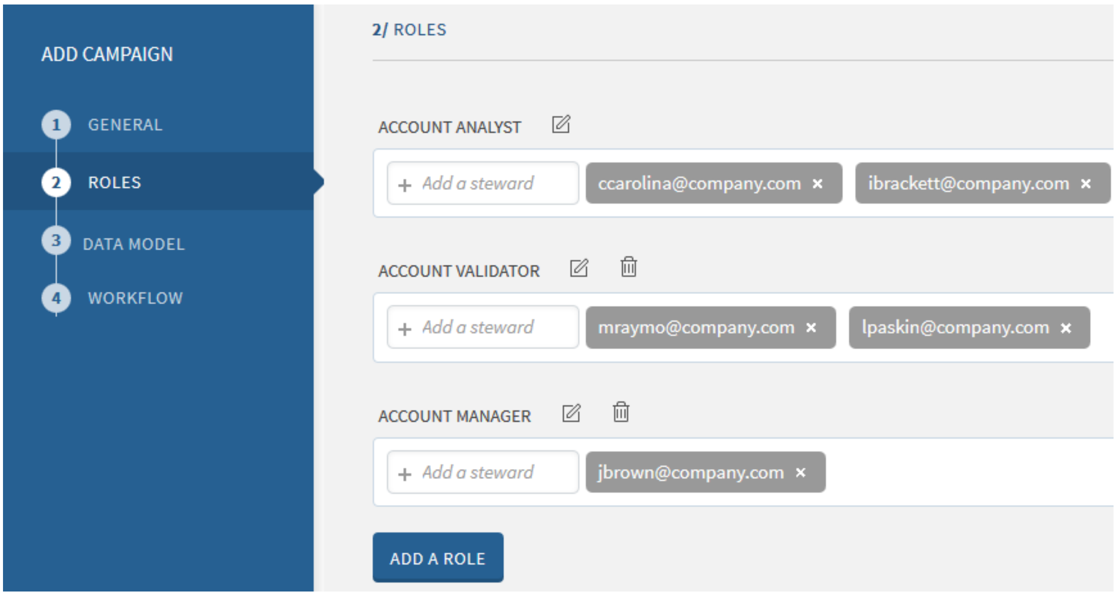
   The due date in a **Merging** campaign is only on master records.
8. Click in the **CAMPAIGN OWNERS** field and select from the list one or more users to grant them access on the current campaign.
   This list shows all the users defined in Talend Administration Center and assigned the role of campaign owners.

Defining roles in the campaign

As a campaign owner, you can define roles in campaigns to decide which data stewards are part of the campaign and which data stewards can access what tasks and in what transition of the campaign workflow.

For this `Reconciling client data` campaign, you want to define three roles.

1. In the home page, click **ROLES** > .
2. Name the first role `ACCOUNT ANALYST` and click to save it.



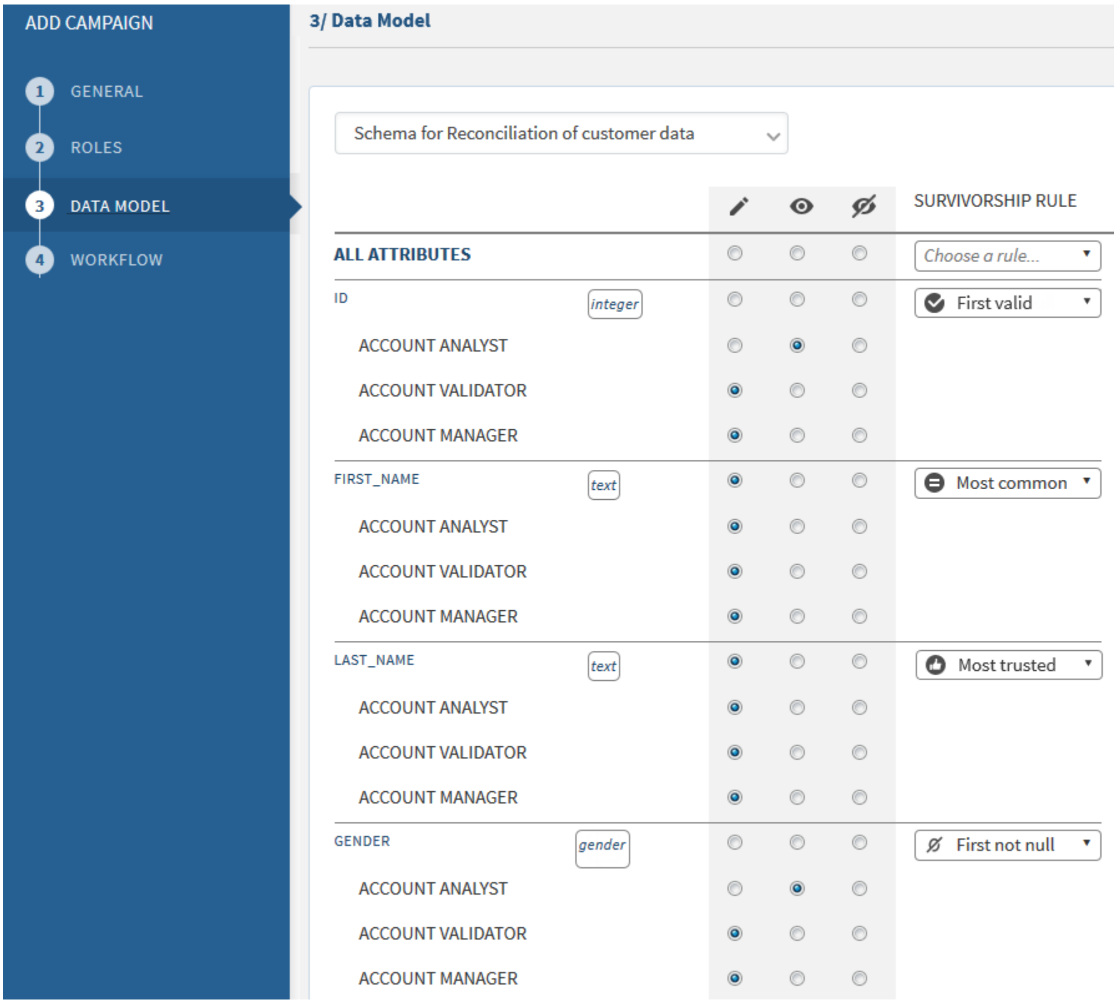
3. Repeat the above steps to add two other roles and name them `ACCOUNT VALIDATOR` and `ACCOUNT MANAGER`.
4. Click in the field and select from the list one or more data stewards to whom you want to assign the role.
   This list shows all the users defined in Talend Administration Center and assigned the role of data stewards.

Setting a data model in the campaign

When defining the campaign, you need to select what data model to use for the syntactic and semantic validation of data and decide the read/write access permission per role to each of the attributes in the selected data model.

The data model used in the campaign decides the structure of the data to be managed.

1. In the home page, click **Model** and select from the model list the data structure you want to use in the campaign.
   In this example, the data model `Schema for Recociliation of customer data` is created for the `Reconciling client data` campaign. The model list gives access to all the data models that have been already defined on the Talend Data Stewardship server.



2. Select the buttons next to each of the attributes in the data structure to set permission per attribute and per data steward and define who can view/edit which attributes.

Option	Description
	gives a read/write access to the attribute in the data model.
	gives only a read access to the attribute in the data model. This type of access is useful if the data steward needs to access the information to make a relevant decision but must not change the value, for instance unique identifiers of other elements linked to the entity the steward is viewing, or data that you know is reliable and must not be changed.
	gives no access to the attribute. Hiding an attribute is useful if the information is sensitive and should not be visible by the data steward, financial information for instance. Another example of attributes to be hidden is if the information is just noise for the steward, technical identifier for instance, but need to be propagated as part of the task.

For example, in this campaign you grant a read-only access to the identifier attribute for the campaign participants who have the `ACCOUNT ANALYST` role. While other participants have a read-write access.

3. Select a rule from the **Survivorship Rule** lists next to each of the attributes.
   These rules are automatically used to decide what attribute values define the master records when loading data into the campaign. Data stewards can then manually modify these choices.

Option	Description
First valid	first source should contains a valid value, where "first" is defined by the order of the records when the task is created.
First not null	first source should contains a value, where "first" is defined by the order of the records when the task is created.
Most common	selects the most common attribute value of the duplicates coming from one or more data sources.
Most recent	selects the most recent attribute value of the duplicates coming from one or more data sources. This is based on the metadata of the last update date.
Most trusted	selects the most trusted attribute value of the duplicates as per the trust score you set when creating the campaign or when loading the tasks in the campaign. If no trust score is defined, this option does not work.

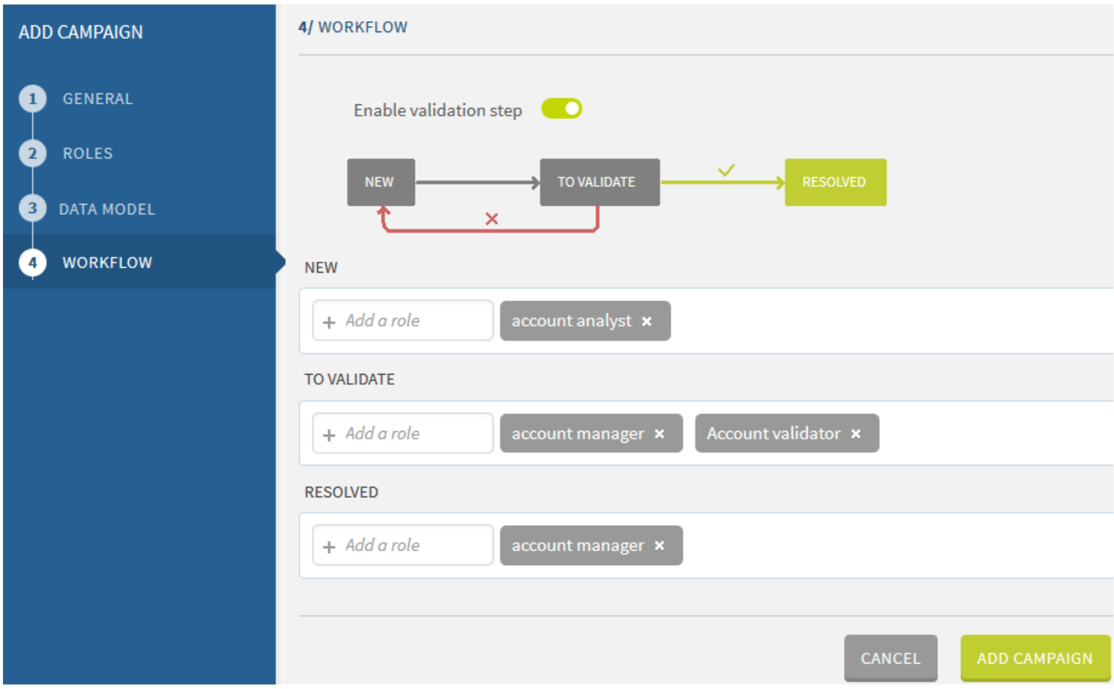
   You can select one rule for all the attributes by selecting it from the list in the top-right corner of the form. If a given algorithm cannot be applied, the rule falls back to **First not null**. For example, if you do not set a trust score and you select **Most trusted** during the campaign definition, **First not null** is used in place. Similarly, **First not null** is used if you select **Most common** or **First valid** and there are no common or no valid values among the data duplicates.

Defining a workflow in the campaign

For the `Reconciling client data` campaign, you want to dd a validation step to the two-state workflow. This enables the steward who is added in the **TO VALIDATE** field to validate the modifications done on the tasks and mark them as resolved.

Workflows control the transitioning of tasks from one state to another. They define how data stewards can collaborate to resolve the campaign tasks, that is which stewards can access which task state or transition.

1. Click the switch next to **Enable validation step** to add a validation step.
2. Click in the **NEW** field and select the roles, from those you defined in the **Role** section, for which you want to give access to the new tasks.



3. Do the same in the **TO VALIDATE** and **RESOLVED** fields and decide what data stewards should have access to the tasks to be validated and to the resolved tasks, respectively.

   Data stewards who are assigned roles in the **RESOLVED** field can access the list of the resolved tasks. They can reopen one or more tasks and decide to send them back to the initial state or to the validation state, if the later is defined in the workflow.

   For further information about reopening tasks, see the online publication on Talend Help Center (https://help.talend.com).
4. Click **ADD CAMPAIGN** to create the campaign and save it on the server.

Use a Talend Job to load tasks into the new campaign.

You can retrieve the Job and the input file from the online version of this page at https://help.talend.com.

Assigning reconciliation tasks

Once the campaign is created and tasks are loaded in it, the campaign owner can assign the tasks to any data steward part of the campaign.

This document also helps you getting familiar with the Talend Data Stewardship interface.

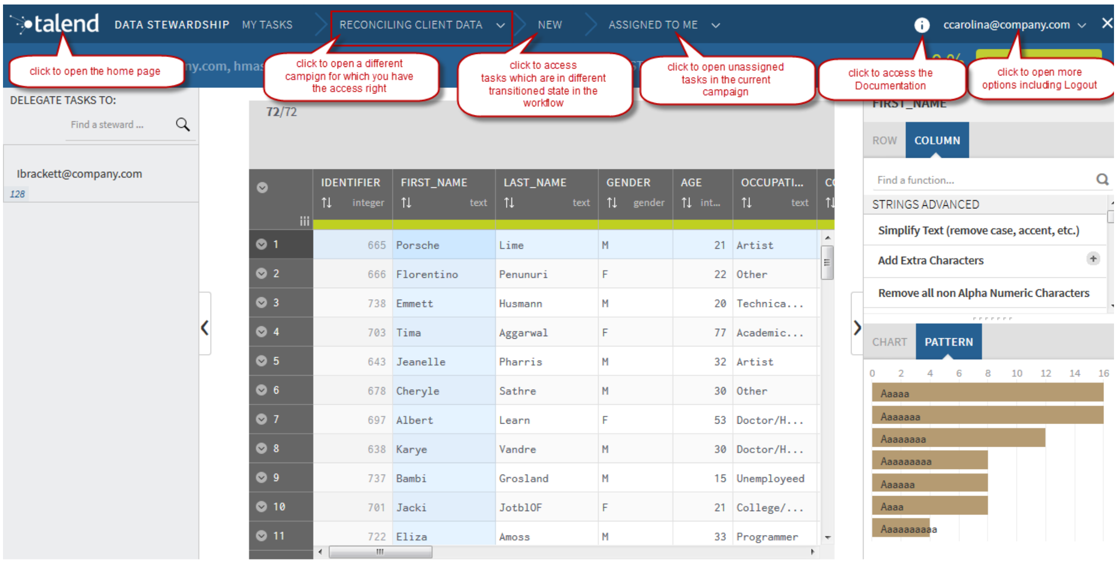
- A campaign owner has defined data stewards in the campaign.
- A campaign owner has loaded the tasks in the campaign using a Talend Job.
- You have accessed Talend Data Stewardship as a campaign owner.

Getting familiar with the Talend Data Stewardship interface

Talend Data Stewardship web application has an easy-to-use interface which make your intervention on data easier than ever.

From the menu bar, you can interact with the content to navigate and access other pages for which you have the right according to the role assigned to you. The capture below shows these actions.

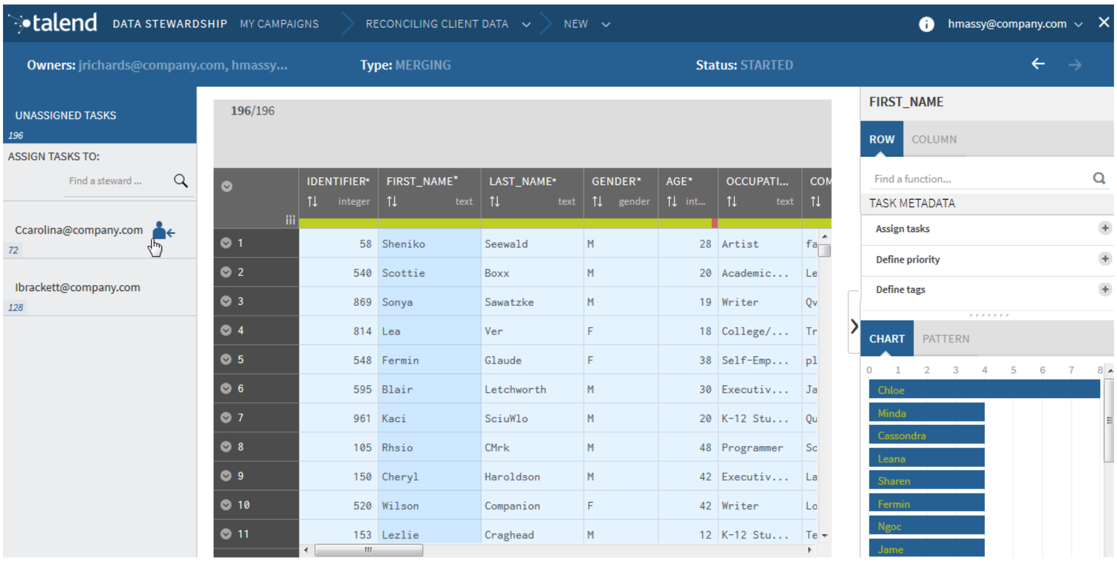
1. Log in as a campaign owner or as a data steward.
2. Click a campaign name to open the task list.

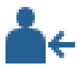


3. Use the menu items to navigate and access other pages for which you have the right according to the role assigned to you.

Assigning tasks

1. Log in as a campaign owner.
2. Select **CAMPAIGNS** and click the `Reconciling client data` campaign.



3. In the list of unassigned tasks, hold the SHIFT key, select the tasks and either:
   - point to the data steward to whom to assign the tasks and click  next to his/her name.
   - Drop the selected tasks to a data steward in the left panel.

   You can use the charts displayed next to the task list to filter data and list only the task you want to assign to a specific data steward.
4. Click **UNASSIGNED TASKS** in the top left corner to go back to the list of unassigned tasks and repeat the above steps to assign tasks to another data steward.

The selected tasks are moved to the list of the data steward. The number of tasks which show under the participant name in the left panel is the number of tasks assigned to him/her in the campaign.

Handling reconciliation tasks

Once the tasks are created in the **Merging** campaign and assigned to you as a data steward, you can start processing duplicate records following the transactions defined in the workflow.

The workflow defined at the campaign creation determines which states are available to what data stewards.

In this example, tasks are transitioned among three states:

- In the first state, values which define the master records are selected based on different scores and potential master records are sent for approval,
- In the second state, potential master records are approved or rejected,
- In the last state, master records are resolved.

Setting rules and values for master records

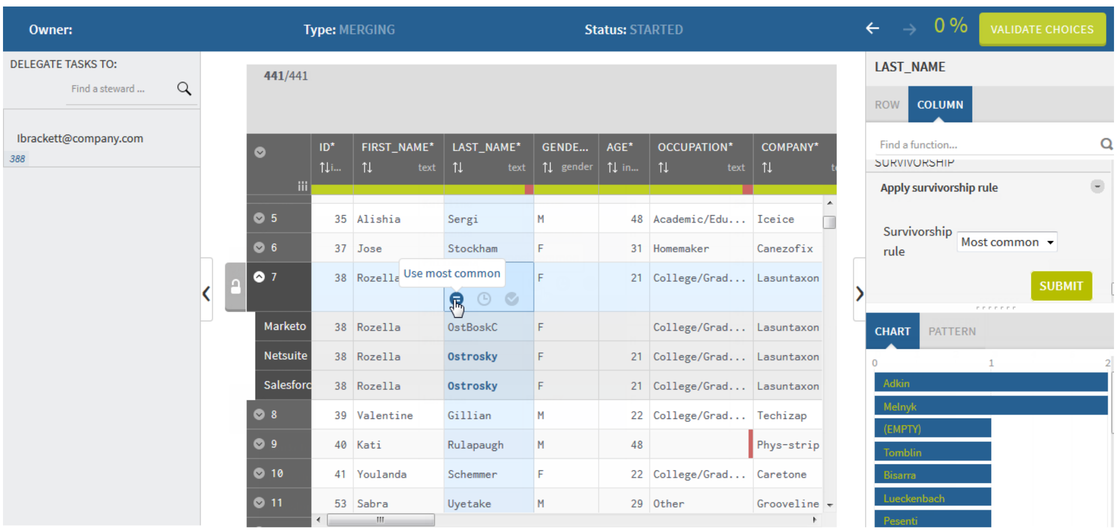
In this example, client duplicate records come from different sources, but Talend Data Stewardship determines initially which attributes of matched records to use to create the master record according to the survivorship rules defined when creating the campaign.

Data stewards can then review the tasks and manually modify survivorship rules per record attribute or enter completely new values to reach the most accurate and reliable master records.

- A campaign owner has created the campaign and granted you access to it.

  For further information, see the online publication on defining roles on Talend Help Center (https://help.talend.com).
- A campaign owner has assigned you tasks in the campaign.
- You have accessed Talend Data Stewardship as a data steward.

1. In the **MY TASKS** page, click the campaign name, `RECONCILING CIENT DATA` in this example, to open a list of the tasks assigned to you.
   The quality bar at the top of the list uses colors to give you a clear view about the quality of the data in each of the columns. Pointing to a color gives you details about the data values in the selected column.



2. To filter the data on which you want to work, click a color in the quality bar on top of a column to list the tasks which match the color indication:

Option	Description
Green	represents valid data which matches the columns type.
White	represents empty fields. However, an empty value for a mandatory field is marked as red, not white.
Red	represents invalid data which does not match the column type or the parameter set in the data model.

3. Click the down arrow on the top-left corner of the task list to expand all the tasks, or click the down arrow of a specific task to expand it.
4. Set survivorship rules to select attributes from customer records and use them to build the master records. Several approaches are possible:
    - Set a survivorship rule manually for one or several attributes of a record: point to an attribute in the master record of a task and from the icons which display, select the survivorship rule you want to apply.
        - : selects the first valid attribute value among the duplicates. "First" is defined by the order of the records when the task is created.
        - : selects the most common attribute value among the duplicates.
        - : selects the most recent attribute value among the duplicates.
        - : selects the most trusted attribute value among the duplicates.

        Survivorship icons are grayed out when the survivorship rule is not applicable on the selected record.
    - Set a survivorship rule manually for one attribute of multiple records.
        1. Click a column heading, `Last_Name` for example, and in the right-hand panel browse to the **Survivorship** section.
        2. Click the button and from the **Survivorship rule** list, select `Most common` as the survivorship rule you want to apply to the name attribute in all the tasks in the list.
        3. Click **Submit** to select the most common name values and add them to the master records of all the tasks.
    - Select the value of a given source attribute to be the value for the master record: point to a source attribute and click the up arrow to set the selected value in the master record.
5. Double-click the value in the master record and set a value of your choice which is not present in any of the sources.
6. If the lock icon has a red background color, correct the invalid value in the task before you can mark it as ready to be validated.
7. Repeat the above step to merge records and create master records for all the tasks assigned to you.
8. Click the icon next to the data record you modified to mark the task as ready to be validated. The first field is marked with green background and a percentage of the completion of your tasks is calculated and displayed in the top right corner.

    You can remodify the records ready to be validated, but this puts the task back to its initial state with a dark-grey background color. You need to reclick the lock icon to mark the task as ready for validation.
9. Click **VALIDATE CHOICES** in the top right corner to validate the changes and move the task from your list.

Master records are created and the records which are validated are moved to the list of the campaign participant who is granted the `ACCOUNT VALIDATOR` role in this example.

Approving changes in master records

In the `RECONCILING CLIENT DATA` campaign created for this example, master records which are created and validated are transitioned to the second step in the workflow. The campaign participant who is granted the `ACCOUNT VALIDATOR` role should now approve or reject the choices done on the master records.

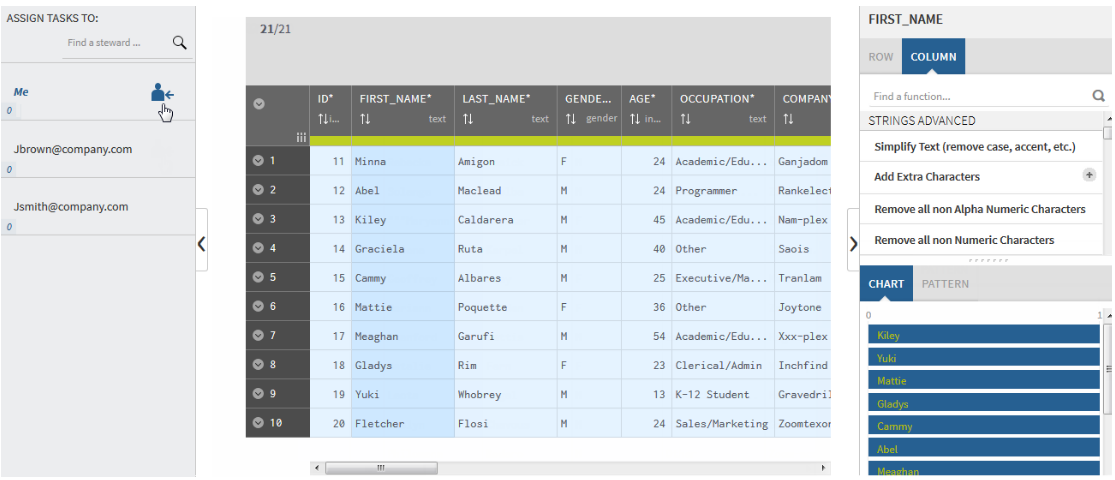
When more than one data stewards are granted the same role in the campaign, the tasks which arrive on their dashboards are unassigned. They should first assign the tasks among themselves before they can start working on them.

- A campaign owner has created the campaign and granted you access to it.

  For further information, see the online publication on defining roles on Talend Help Center (https://help.talend.com).
- Master records are created and validated in the campaign.

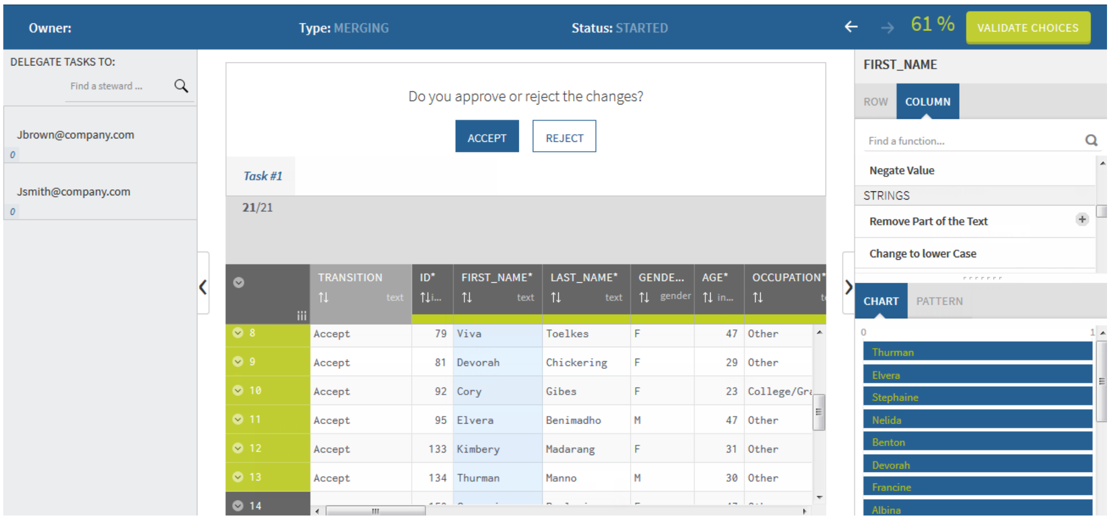
  For further information, see Setting rules and values for master records on page 15.
- You have accessed Talend Data Stewardship as a data steward.

1. In the **MY TASKS** page, click the campaign name, `RECONCILING CIIENT DATA`, to open a list of the tasks transitioned to you.



2. In the list of unassigned tasks, hold the SHIFT key, select the tasks and either:
   - point to the data steward to whom to assign the tasks and click  next to his/her name, or
   - drop them to a data steward in the left panel. Assign them to yourself in this example.

   The selected tasks are moved to your list. The number of tasks which show under your name in the left panel is the number of tasks assigned to you in the current campaign.
   You can use the charts displayed next to the task list to filter data and list only the task you want to assign to a specific data steward.
3. On the menu bar, click **UNASSIGNED** and select **Assigned to me** to display the tasks for which you need to validate the choices of the initial data stewards.



4. Approve the changes on the master records:
   a) to accept the changes on the master records without any more modifications, select the task(s) and click **ACCEPT**.
   b) to accept the changes on a master record after doing some more modifications, use any of the functions under **COLUMN** in the right panel or enter completely new values in the fields and click **ACCEPT**.

   The first fields of the approved task(s) are marked with green background and the **TRANSITION** column is filled in with the decision text, **Accept** or **Reject**.
5. To reject the changes done on the master records, select the records and click **REJECT**.
6. Click **VALIDATE CHOICES** in the top right corner.

The approved tasks are resolved and moved from the list. The data steward who has the `ACCOUNT MANAGER` role in this example can access the resolved tasks and decide to reopen one or more of the tasks, if needed. The rejected tasks are moved back to the initial list of unassigned tasks.

Reopening resolved tasks

Reopening the tasks that have been resolved and transitioned to the final state in the workflow is sometimes necessary to fix the updates done on the values in master records.

In the `RECONCILING CLIENT DATA` campaign created for this example, master records are approved and marked as resolved. The data steward who is granted the `ACCOUNT MANAGER` role can now access the list of the resolved tasks and decide whether it is necessary to reopen some given tasks and fix the updates done on the values in master records.

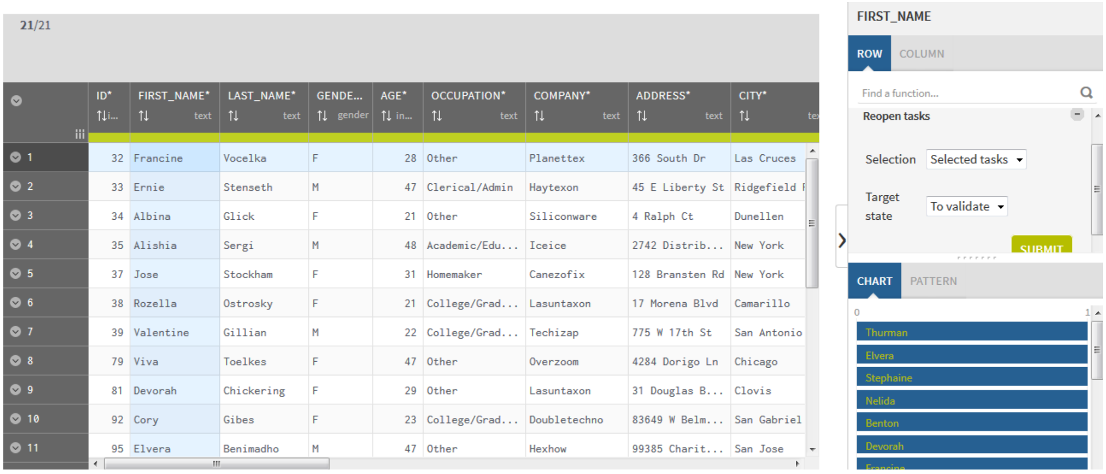
- A campaign owner has created the campaign and granted you access to it.

  For further information, see the online publication on defining roles on Talend Help Center (https://help.talend.com).
- Master records are approved and transitioned to the last step in the workflow.

  For further information, see Approving changes in master records on page 17.
- You have accessed Talend Data Stewardship as a data steward.

1. In the **MY TASKS** page, click the campaign name, `RECONCILING CIIENT DATA`, to open a list of the resolved tasks.

Resolved tasks are always unassigned. However, the data steward who has the `ACCOUNT MANAGER` role in this example can access them.



2. Select a task or hold the SHIFT/Ctrl key and select the tasks for which you need to take a different decision.
3. In the right panel, click **Row** and:
   a) Expand **Reopen tasks**.
   b) Select **Selected tasks** from the **Selection** list.
   c) Select **To validate** from the **Target state** list.
   d) Click **SUBMIT**.

The selected tasks are reopened and sent back to the validation state in the workflow.

What is next?

You have learned how to use a **Merging** campaign in Talend Data Stewardship to ensure the consistency of data that has been loaded into the web application.

You have seen how to pick, based on different scores, the best attribute values from the data sources and presents the most accurate and reliable master records for consumption by users and systems.

To learn about all what you can do with Talend Data Stewardship, see the online documentation on Talend Help Center (https://help.talend.com).

To learn more about Talend different products and solutions, visit https://www.talend.com.