Talend Data Stewardship Use Cases

6.4.1

Contents

Copyright

Adapted for 6.4.1. Supersedes previous releases.

Publication date: June 29th, 2017

Use cases of Talend Data Stewardship

The use cases described for Talend Data Stewardship show several means of organizing the interactions on data whenever human intervention is required.

Consider the below scenarios:

- Beta testing: your enterprise has a beta testing project. Beta testing begins with the selection of test candidates. An **ARBITRATION** campaign enables you to select the appropriate number and composition of participants.

  For further information, see Adding an Arbitration campaign for Beta testing on page 4 and Handling arbitration tasks to select Beta testers on page 14.
- Approving data in a product catalog: management of product information includes referencing product and making sure it holds the most complete and accurate information. Suppose that you have a catalog with product records where some values need to be fixed or provided. A **RESOLUTION** campaign enables you to cure such records.

  For detail, see Adding a Resolution campaign to curate product records on page 15 and Handling resolution tasks to curate product records on page 26.
- Managing contacts: the sales team has found a number of duplicates in the Contacts objects in the enterprise Customer Relationship Management system (CRM) and this hurts their productivity. A **MERGING** campaign enables you to solve the duplicates by surviving only the appropriate data.

  For further information, see Adding a Merging campaign to deduplicate records on page 28 and Handling merging tasks to deduplicate records on page 38.
- Matching pairs of records: suppose that you have a million of pairs of suspect records for which you need to identify if the second element in each pair is a duplicate. Such process can be a pain point if you do not have the right tool to do this. A **GROUPING** campaign brings you the answer here.

  For further information, see Adding a Grouping campaign to identify duplicate pairs on page 40 and Handling grouping tasks to decide on relationship among pairs of records on page 51.

Adding an Arbitration campaign for Beta testing

An **Arbitration** campaign defines a list of possible arbitration choices for the campaign tasks. The outcome of an arbitration task is the arbitration choice made by data stewards.

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

Consider your enterprise has a beta testing project for one of its products. Beta testing begins with the selection of the test candidates, usually from existing customers because they match the product target market.

- 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.

To add the campaign, you need to accomplish the below procedures:

1. **** Defining a data model for the Arbitration Campaign on page 5.
2. Defining the Arbitration campaign on page 6.
3. Defining roles in the Arbitration campaign on page 8.
4. Setting a data model in the Arbitration Campaign on page 8.

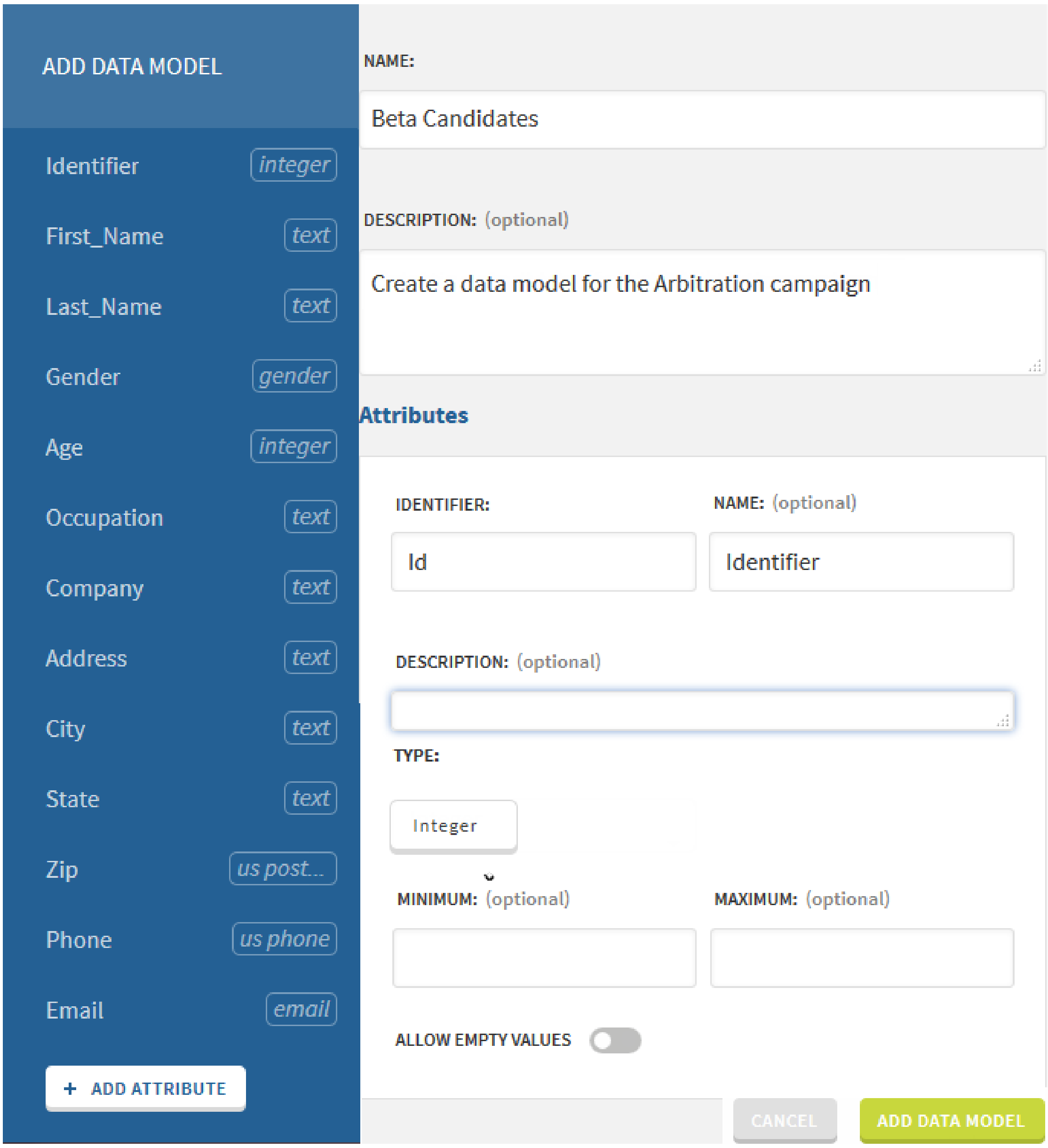
5. Defining a workflow in the Arbitration campaign on page 10.

Defining a data model for the Arbitration Campaign

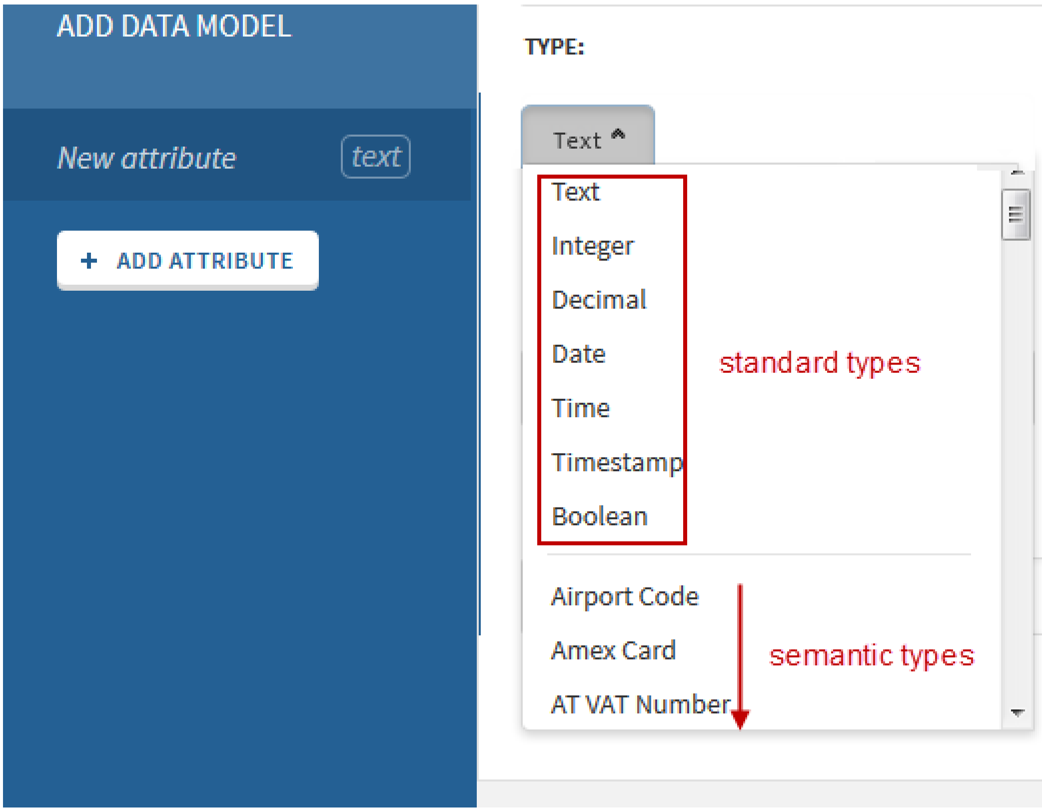
In this example, you create a data model to determine the structure of the data to be managed in the `Beta Candidates` campaign which you create to allow data stewards to select the appropriate candidates of a software beta testers.

Talend Data Stewardship has data model awareness which makes possible the syntactic and semantic validation of data. You can define the attributes in the data model and select their types out of a predefined standard or semantic types.

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.



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.

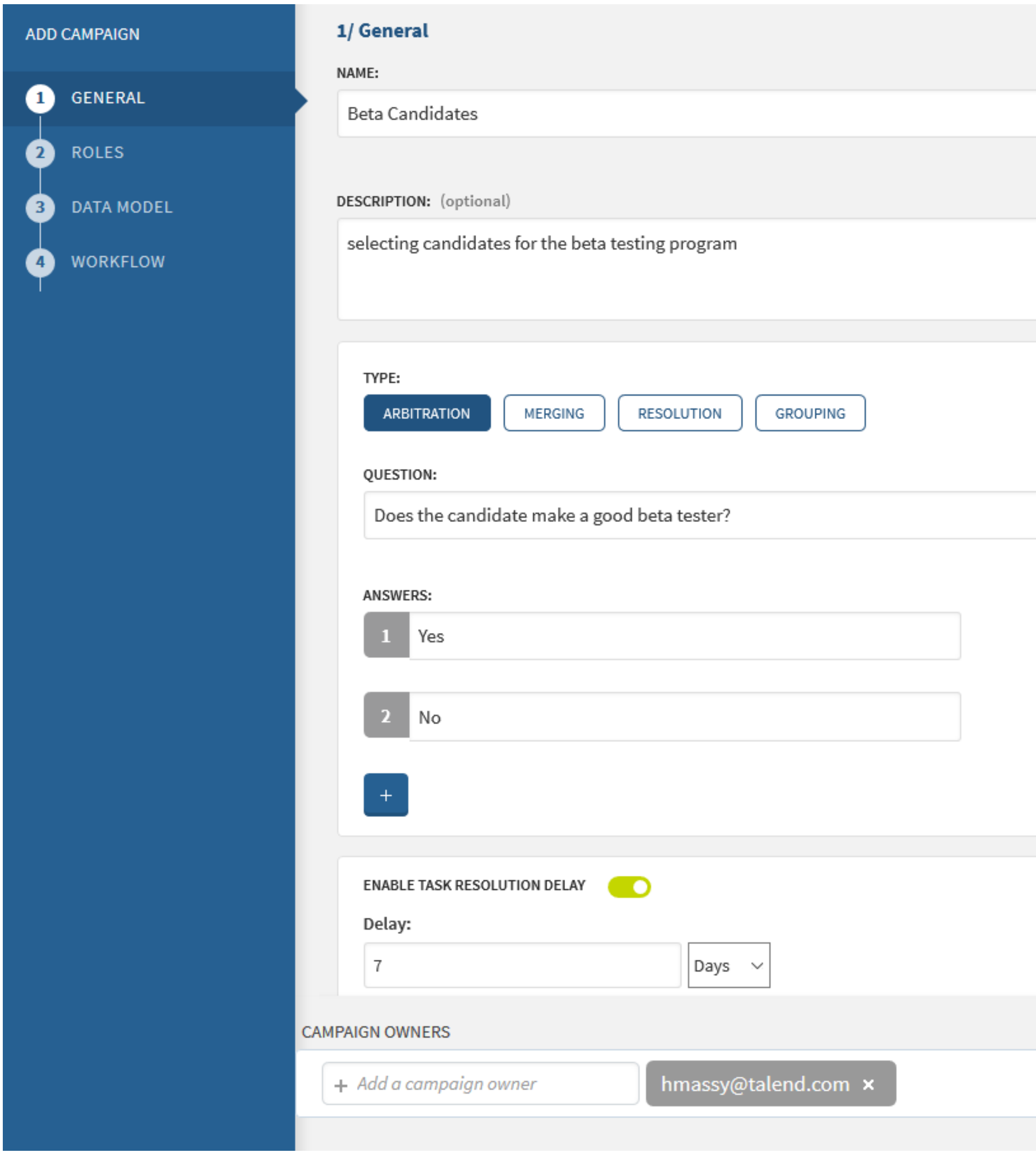
4. If needed, click the switch next to **ALLOW EMPTY VALUES** to disable the load of empty fields to Talend Data Stewardship. This option is enabled by default.
5. Click **ADD ATTRIBUTE** in the left panel and repeat the above steps to create all the columns you need in the data model.
   The columns defined for the `Beta Candidates` campaign used as an example include information about the first and last names of the participants, their occupation and the company in which they work as shown in the capture.

Defining the Arbitration campaign

In this scenario, you define the `Beta Candidates` campaign to allow data stewards to select the appropriate number and composition of participants in a beta testing program.

Start by defining the campaign metadata.

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



4. In the **QUESTION** field, enter the question to use in the campaign, for example: `Does the candidate make a good beta tester?`
5. Enter `Yes` and `No` as the arbitration answers in the two by-default **ANSWERS** fields.
6. Use the + button to add more answer choices when needed.
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.
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.

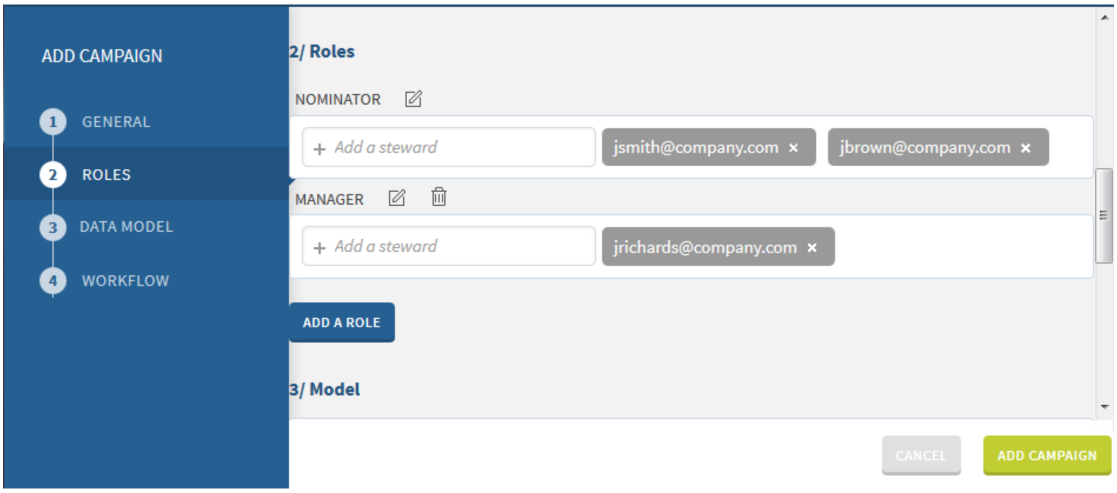
Defining roles in the Arbitration campaign

You need now to define roles in the campaign to decide the actions to be taken on data, which data stewards are part of the campaign and which data stewards can access which tasks and in what transition of the campaign workflow.

You can create and start a campaign without defining any roles and edit the campaign later to add the roles of your choice.

After defining the metadata of the campaign, you need to group data stewards into roles and assign them access rights.

1. In the home page, click **ROLES** > .
2. Name the role `NOMINATOR` and click to save it.
   For this example only one role is needed as the workflow in the campaign will not have a validation step, but you can add as many roles to a campaign as needed.



3. Click the field below the new role and select from the list the steward(s) who will decide what candidates to accept for the beta testing.

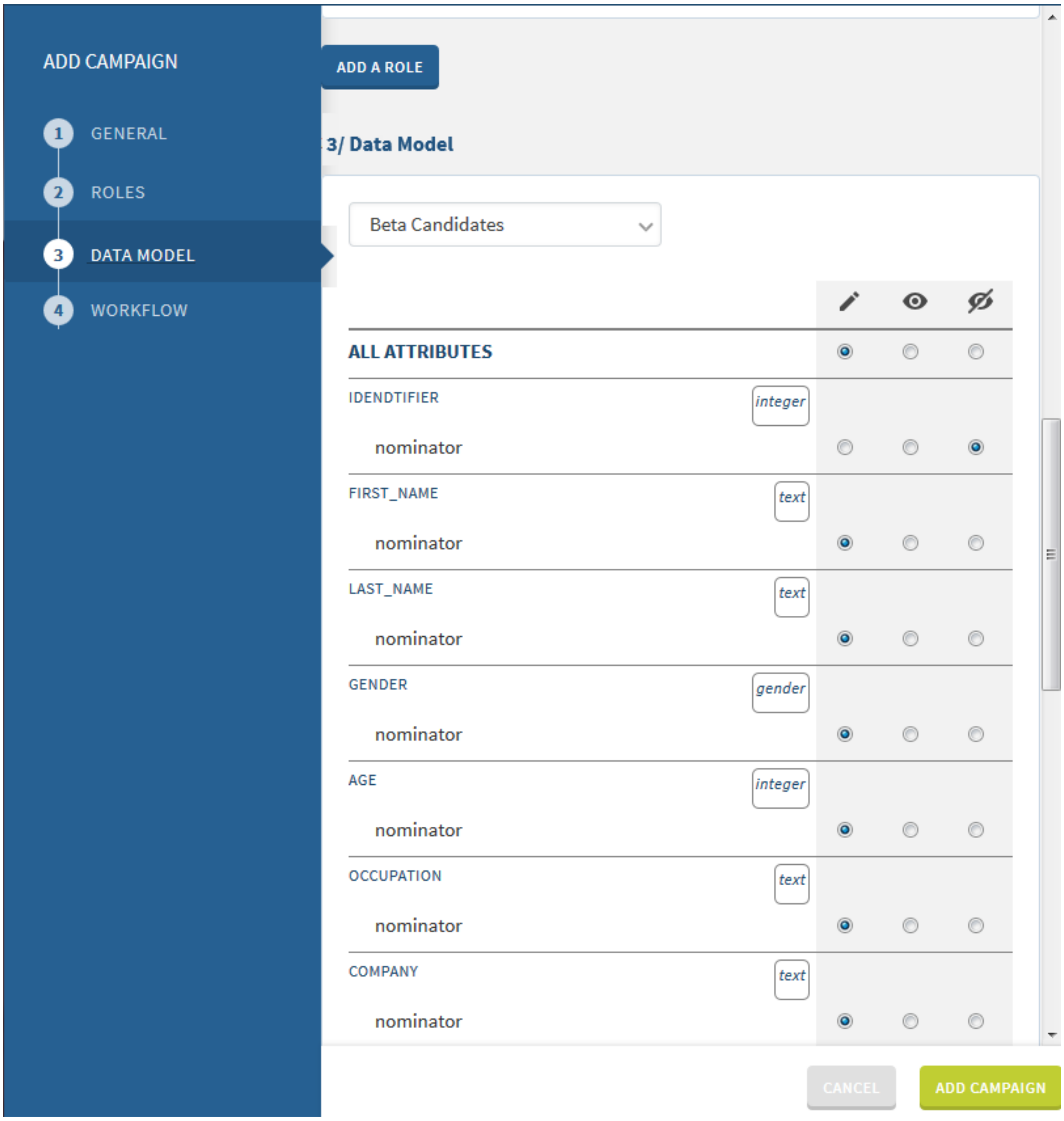
Setting a data model in the Arbitration Campaign

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 **Data Model** and select from the model list the data structure you want to use in the `Beta Candidates` campaign.

   The **Data 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.

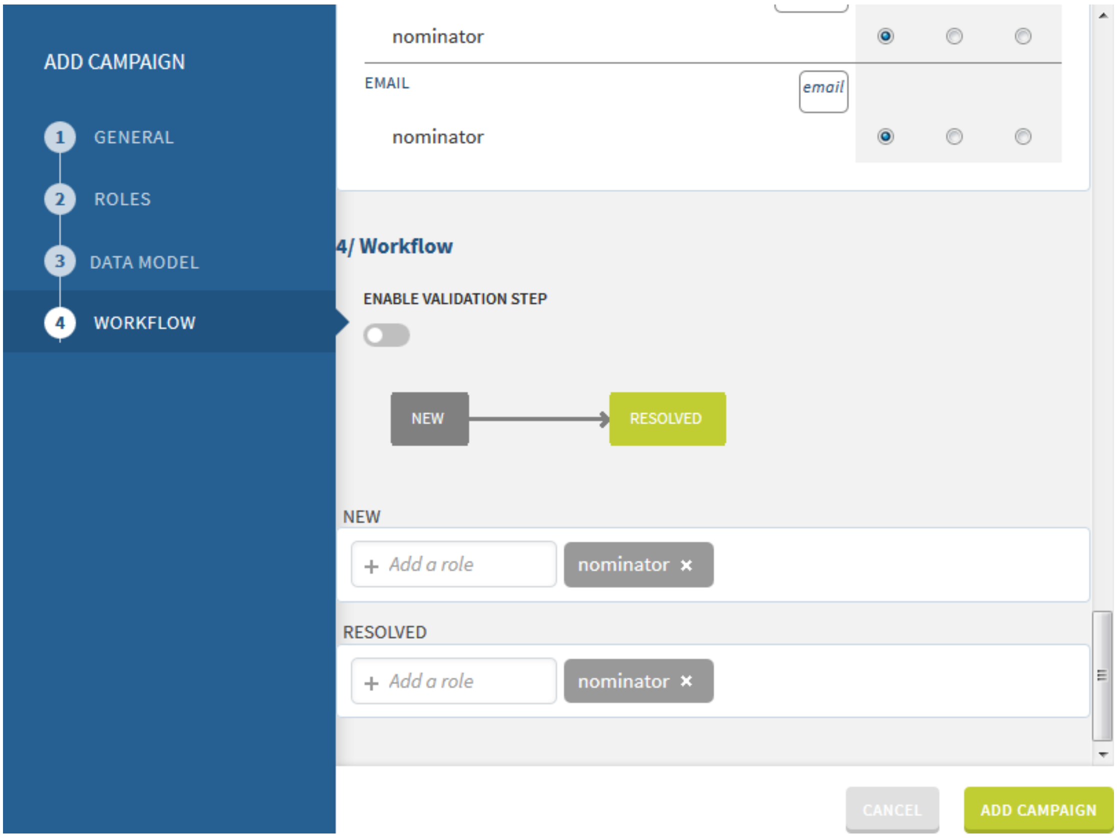
In this example you have only the NOMINATOR role. You want to make the technical identifier invisible to the stewards who are assigned the role and grant them a read/write access to all other attributes.

Defining a workflow in the Arbitration campaign

For the Beta Candidates campaign, you want to use the by-default workflow which has one initial state and one final state. The newly created tasks are assigned to specific data stewards, handled and then marked 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 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.



2. Do the same in the **RESOLVED** field and select the roles which you want to grant access to the resolved tasks.

   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).
3. Click **ADD CAMPAIGN** to create the campaign and save it on the server.

You can now:

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

For further information about writing tasks in campaigns, see the online publications on Talend Help Center (https://help.talend.com/).

2. Assign tasks to data stewardships. For further information, see Assigning and reassigning tasks as a campaign owner on page 11.
3. Start tasks resolution.

   For further information, see Handling arbitration tasks to select Beta testers on page 14.

Assigning and handling arbitration tasks

Assigning and reassigning tasks as a campaign owner

Campaign owners can use the Job properties in Talend Studio to automatically assigned the tasks to a data steward. Otherwise, once logged to Talend Data Stewardship they can assign or reassign the tasks in a specific campaign to any data steward part of the campaign.

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

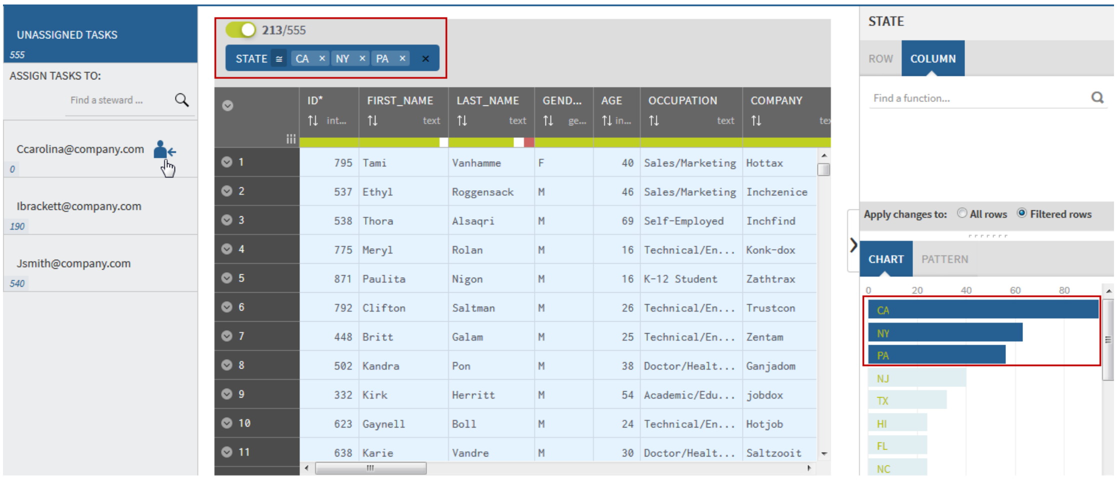
Assigning tasks

This procedure shows how to assign tasks manually.

For further information about assigning multiple tasks using a bulk operation, see the online publication on Talend Help Center (https://help.talend.com).

1. In the home page, select **CAMPAIGNS** and click the name of the campaign which holds the tasks you want to assign.
2. In the list of unassigned tasks, select a task or hold the Ctrl/SHIFT key and click to select multiple tasks and:
   - Drop the selected tasks to a data steward in the left panel, or
   - Point to the data steward to whom to assign the tasks and click next to his/her name.
   - Use the **ASSIGN TASKS TO** field in the top left corner to search for the data steward to whom to assign the tasks and click , or

   You can use the data quality bar on top of the columns or the charts displayed in the right-hand panel to filter data and list only the task you want to assign to specific data stewards.



The selected tasks are moved to the task 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 current workflow state.

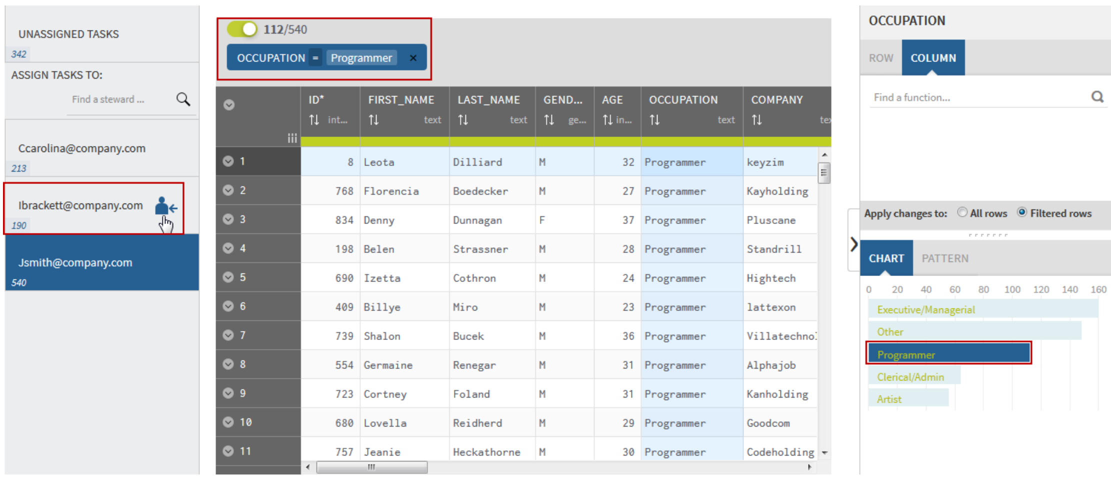
3. In the panel to the left, click **UNASSIGNED TASKS** in the top left corner to go back to the list of unassigned tasks and repeat the above steps to assign tasks to another data steward

Reassigning tasks

Campaign owners can reassign tasks between the data stewards part of the campaign.

1. In the home page, select **CAMPAIGNS** and click the name of the campaign which holds the tasks you want to reassign.
2. In the left-hand panel, click a data steward to display the list of all the tasks assigned to him/her.
3. Select one or multiple tasks from the list and in the left-hand panel:
   - Point to the data steward to whom to reassign the tasks and click  next to his/her name, or
   - Use the **Assign Tasks To** field to search for the data steward to whom you want to reassign the tasks and click .

   You can use the data quality bar on top of the columns or the charts displayed in the right-hand panel to filter data and list only the tasks you want to reassign.



The number of tasks which show under the user name in the left panel is the number of tasks assigned to the user within the campaign.

The selected tasks are moved from the initial data steward to the new data steward.

Assigning and delegating tasks as a data steward

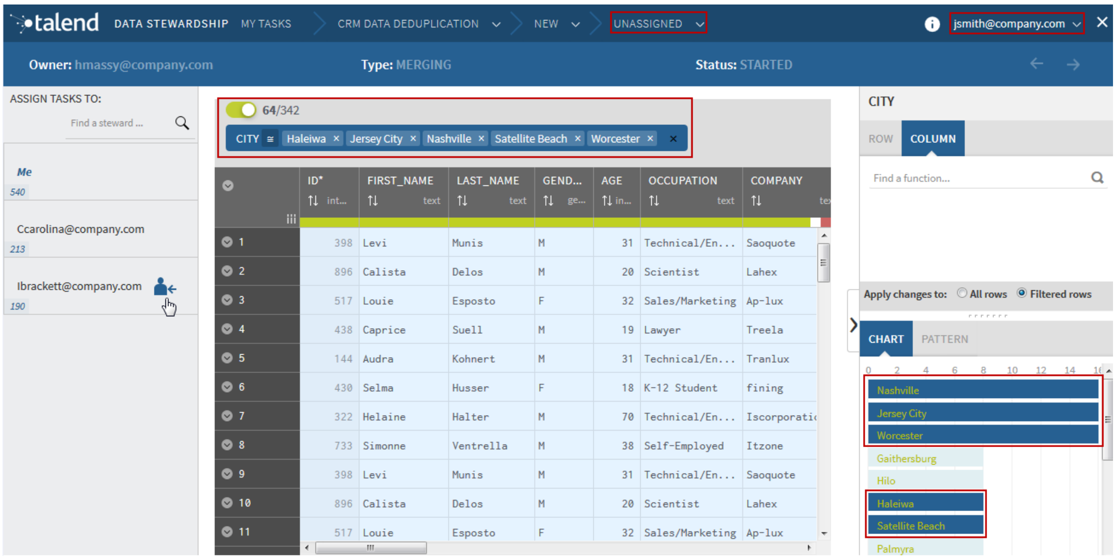
Data stewards can assign new tasks to themselves or other data stewards part of the same campaign. They can assign tasks to other data stewards only if tasks are currently unassigned or assigned to themselves.

Data stewards can assign or delegate tasks to other data stewards to meet certain needs, for example to improve the distribution of workload across business users or because of expertise on a specific task.

- A campaign owner has loaded the tasks in the campaign using a Talend Job.
- A campaign owner has created the campaign and granted you access to it.
- You have accessed Talend Data Stewardship as a data steward.

Assigning tasks

1. In the **TASKS** page, point to the campaign which holds the tasks you want to assign and click the icon that displays to the right of the campaign name.
   A list of all the unassigned tasks in the campaign is opened.



   You can click **Unassigned** on the menu bar and select **ASSIGNED TO ME** to open the list of the task already assigned to you in the campaign.
2. Use the data quality bar on top of the columns or the charts displayed in the right-hand panel to filter data and list only the tasks you want to assign, or select the tasks manually.

   For further information about filtering tasks with the quality bar or the charts, see the online publication on Talend Help Center (https://help.talend.com).
3. In the left-hand panel, either:

   - Point to the data steward to whom to assign the tasks and click  next to his/her name, or

- Use the **Assign Tasks To** field to search for the data steward to whom you want to assign the tasks and click .

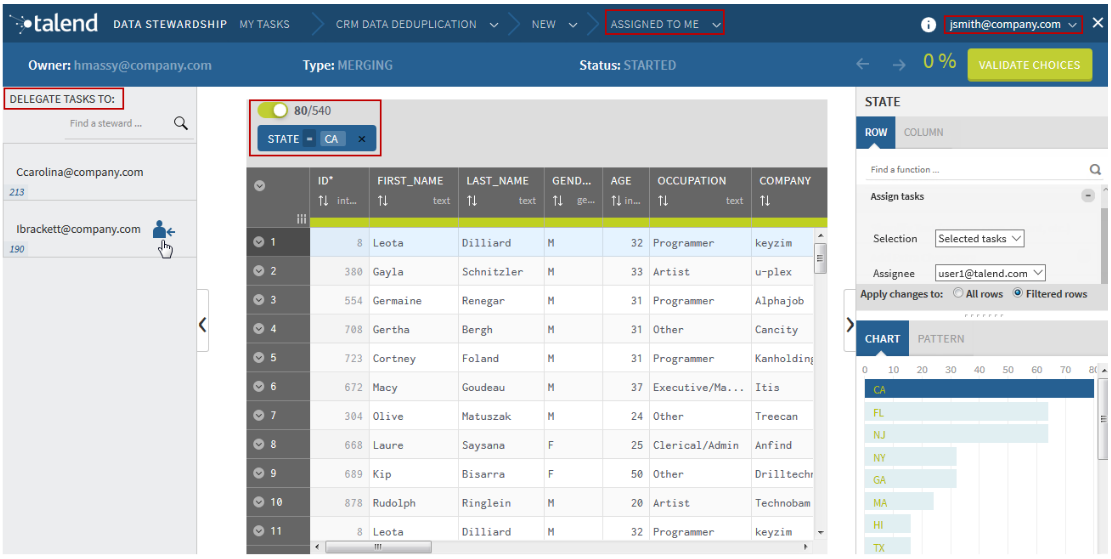
The selected tasks are moved to the dashboard of the data steward. The number of tasks which show under the participant name in the left panel is the number of tasks assigned to him/her within the campaign.

Delegating tasks

The reasons to delegate tasks include the need of certain expertise to handle the tasks, a risk of failing to do the tasks by the appointed time or fast approaching vacations.

Tasks have been assigned to the data stewards.

1. In the **TASKS** page, click the campaign which holds the tasks you want to delegate.



2. Use the chart in the right-hand panel to filter and select a group of the tasks assigned to you and in the left-hand panel, and do one of the following:
   - In the panel to the right, click **ROW** > **Assign tasks** and use the two lists to decide what tasks to delegate to what data steward.
   - Point to the data steward to whom to delegate the tasks and click  next to his/her name.
   - Use the **Delegate Tasks To** field to search for the data steward to whom you want to assign the tasks and click  next to his/her name.

Handling arbitration tasks to select Beta testers

Arbitration tasks consist in choosing among a list of arbitration choices one choice for each data record. Once you validate your choice, you transition the task to the second state defined in the workflow. The workflow defined at the campaign creation determines which states are available to what data stewards.

In this example, you need to select from a customer list the appropriate candidates for a beta test planned by your enterprise. Once the list is created, all the processes of a beta program is made much easier, for example distributing the product, collecting and evaluating feedback and so on.

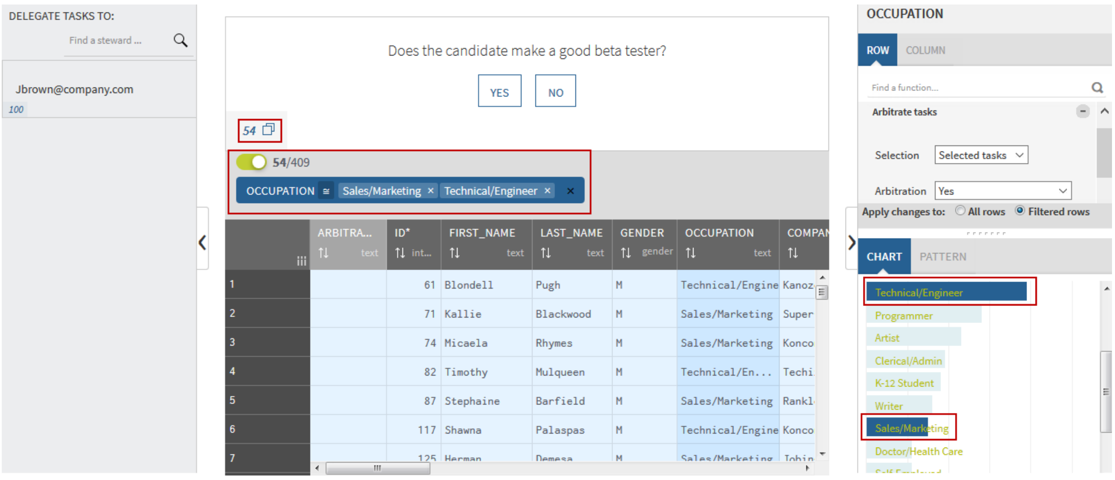
- A campaign owner has created the campaign and granted you access to it.

For further information, see Adding an Arbitration campaign for Beta testing on page 4.

- A campaign owner has assigned you tasks in the campaign. Otherwise, you can assign the tasks to yourself.

  For further information, see the online publication about assigning tasks on Talend Help Center (https://help.talend.com).

1. In the **TASKS** page, click the campaign name, `Beta Candidates` in this example, to open a list of the tasks assigned to you.



2. Select one task and click a choice from the list, or use the right-hand panel and set a choice for multiple tasks at once.

   In this example, you want to accept for beta testing all the candidates from the marketing and technical departments. In such a case, you use the chart in the right-hand panel to filter your records by the `Occupation` column before you select the tasks and click on `YES` for your choice.

   For further information about selecting arbitration choices using a bulk operation, see the online publication on Talend Help Center (https://help.talend.com).

   Listed tasks are tagged with green to show that a decision has been taken on them.
3. Click **VALIDATE CHOICES** in the top-right corner of the page to validate the choices you have done on the tasks.

Choices are set, data records are validated and moved from your list. Data records are transitioned to the second step in the workflow where they need to be approved by another data steward or where they are already resolved depending on the workflow defined in the campaign.

Adding a Resolution campaign to curate product records

A **Resolution** campaign enables you to cure any anomalies in data records such as invalid or missing values.

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

Consider a scenario where the enterprise product catalog has product records with some values that need to be fixed or provided. In such a case, you need to approve product information to make sure the catalog holds the most complete and accurate information. Once this is done, a Talend Job can upload cleansed data to the Master data management (MDM) repository.

- 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 defined a data model for the campaign in Talend Data Stewardship.
- You have accessed Talend Data Stewardship as a campaign owner.

To add the campaign, you need to accomplish the below procedures:

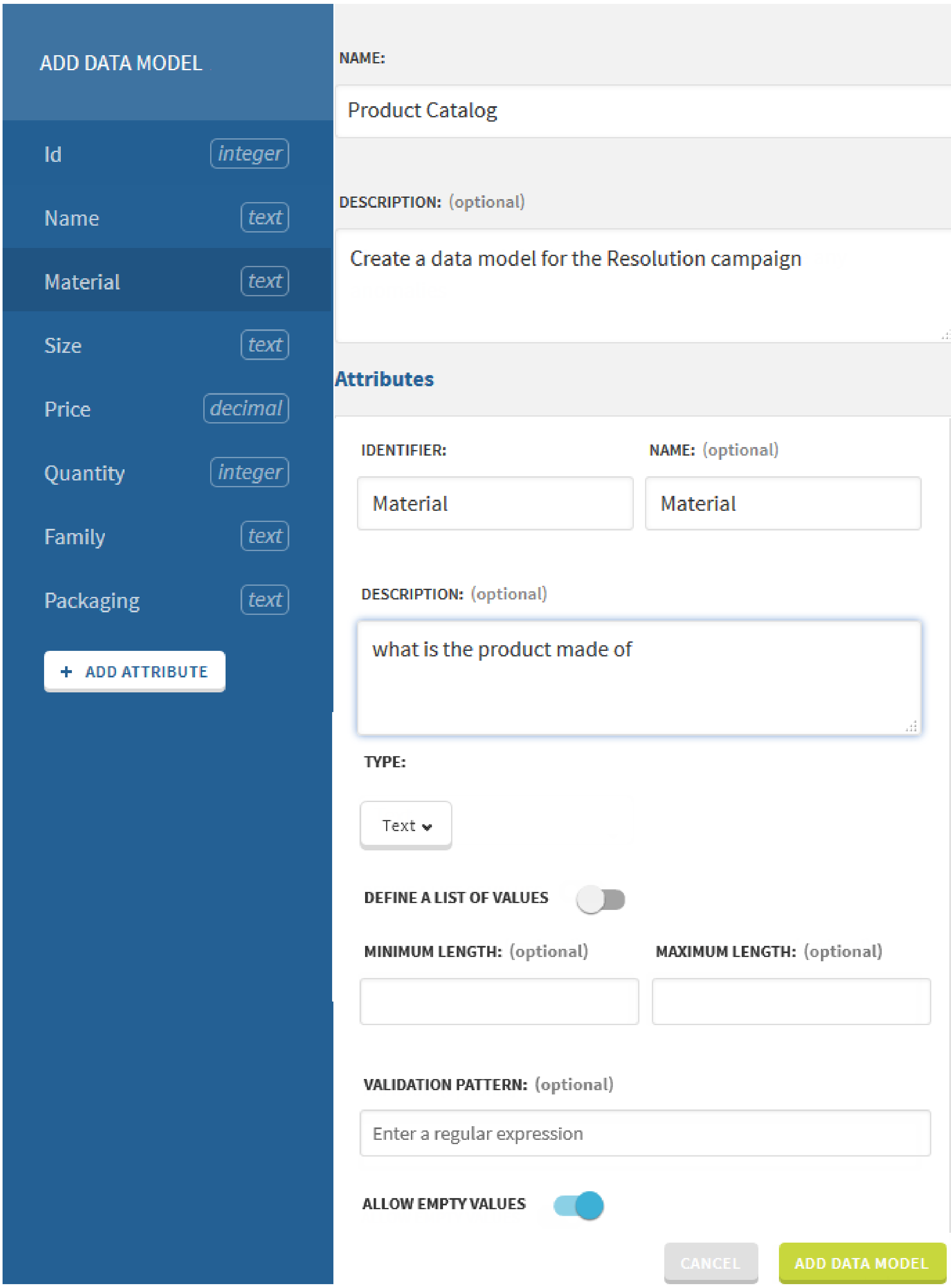
1. Defining a data model for the Resolution campaign on page 16.
2. Defining the Resolution campaign on page 18.
3. Defining roles in the Resolution campaign on page 19.
4. Setting a data model in the Resolution campaign on page 20.
5. Defining a workflow in the Resolution campaign on page 22.

Defining a data model for the Resolution campaign

In this example, you create a data model to determine the structure of the data to be managed in the `Product Catalog` campaign which you create to allow data stewards to curate any anomalies in the data listed in the enterprise catalog.

Talend Data Stewardship has data model awareness which makes possible the syntactic and semantic validation of data. You can define the attributes in the data model and select their types out of a predefined standard or semantic types.

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.

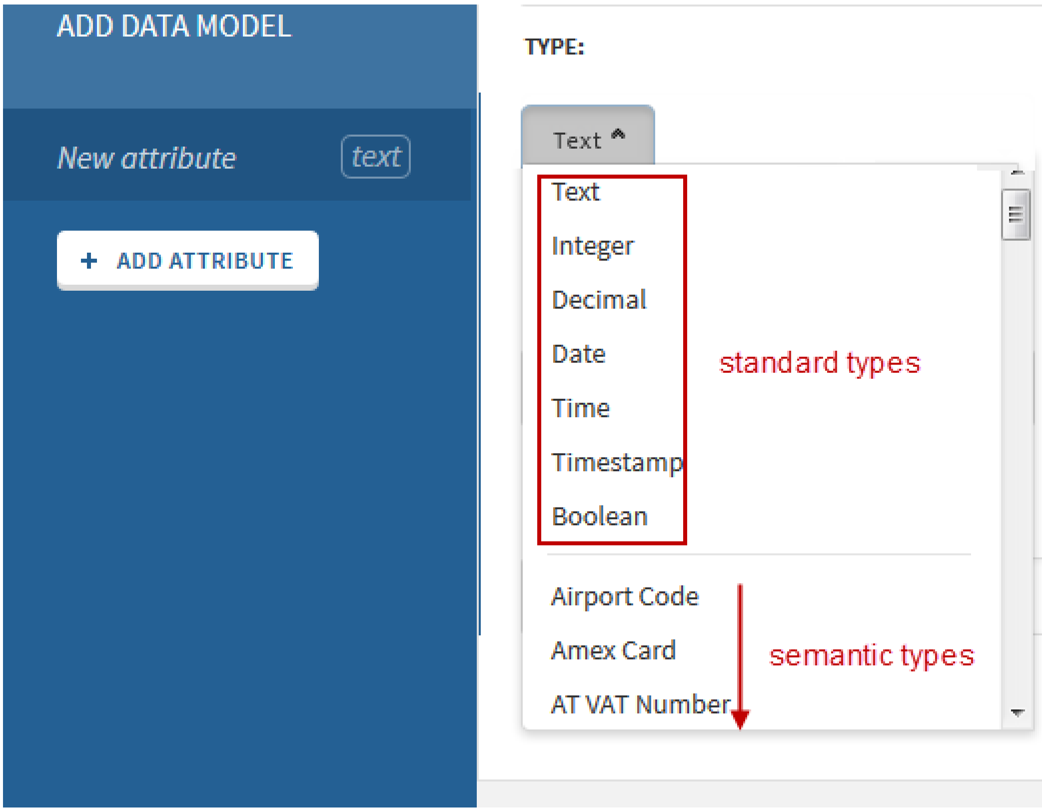


3. In the **Attributes** section, define the columns you want to have in the data model as the following:

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.

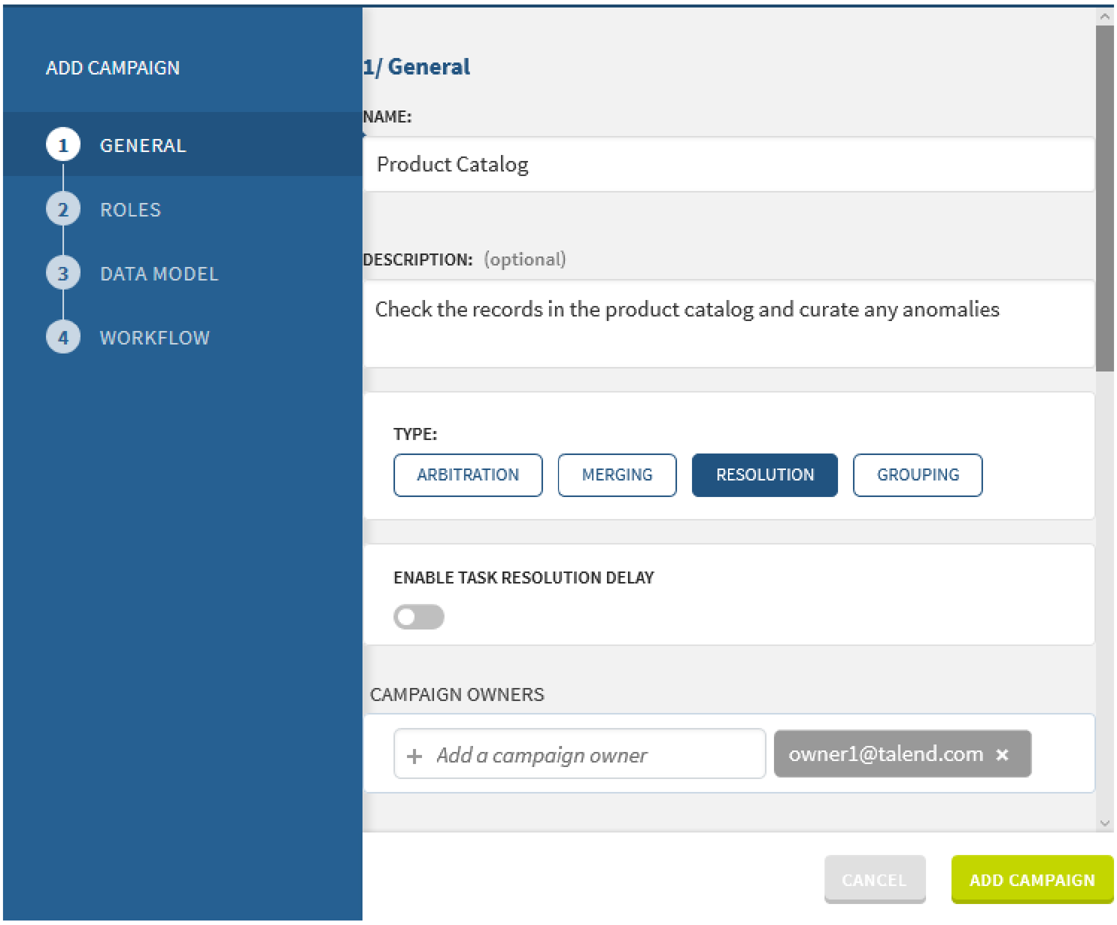
4. If needed, click the switch next to **ALLOW EMPTY VALUES** to disable the load of empty fields to Talend Data Stewardship. This option is enabled by default.
5. Click **ADD ATTRIBUTE** in the left panel and repeat the above steps to create all the columns you need in the data model.
   The columns defined for the `Product Catalog` campaign used as an example include the description of each product item, its size and price as shown in the capture.

Defining the Resolution campaign

In this scenario, you define the `Product Catalog` campaign to allow data stewards to maintain the quality of the product catalog of your enterprise which contains a list of all the products you want to advertise.

Start by defining the campaign metadata.

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



4. 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.
5. 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.

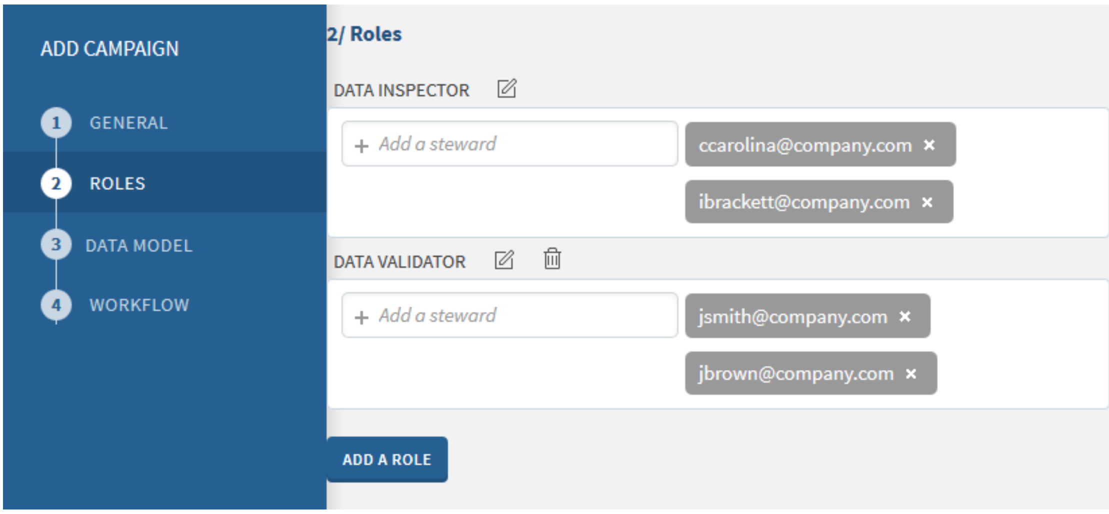
Defining roles in the Resolution campaign

You need now to define roles in the campaign to decide the actions to be taken on data, which data stewards are part of the campaign and which data stewards can access which tasks and in what transition of the campaign workflow.

You can create and start a campaign without defining any roles and edit the campaign later to add the roles of your choice.

After defining the metadata of the campaign, you need to group data stewards into roles and assign them access rights.

1. In the home page, click **ROLES** > .
2. Name the first role `DATA INSPECTOR` and click next to the field to save it.



3. 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.
4. Repeat the above steps to add a second role and name it `DATA VALIDATOR`.
5. From the **Add a steward** list, select the user to validate the choices of the users who are assigned the `DATA INSPECTOR` role.
   In this example, data stewards who have the `DATA INSPECTOR` role inspect data and cure it, and data stewards who have the `DATA VALIDATOR` role validates what has been done by the inspectors. You can add as many roles to the campaign as needed.

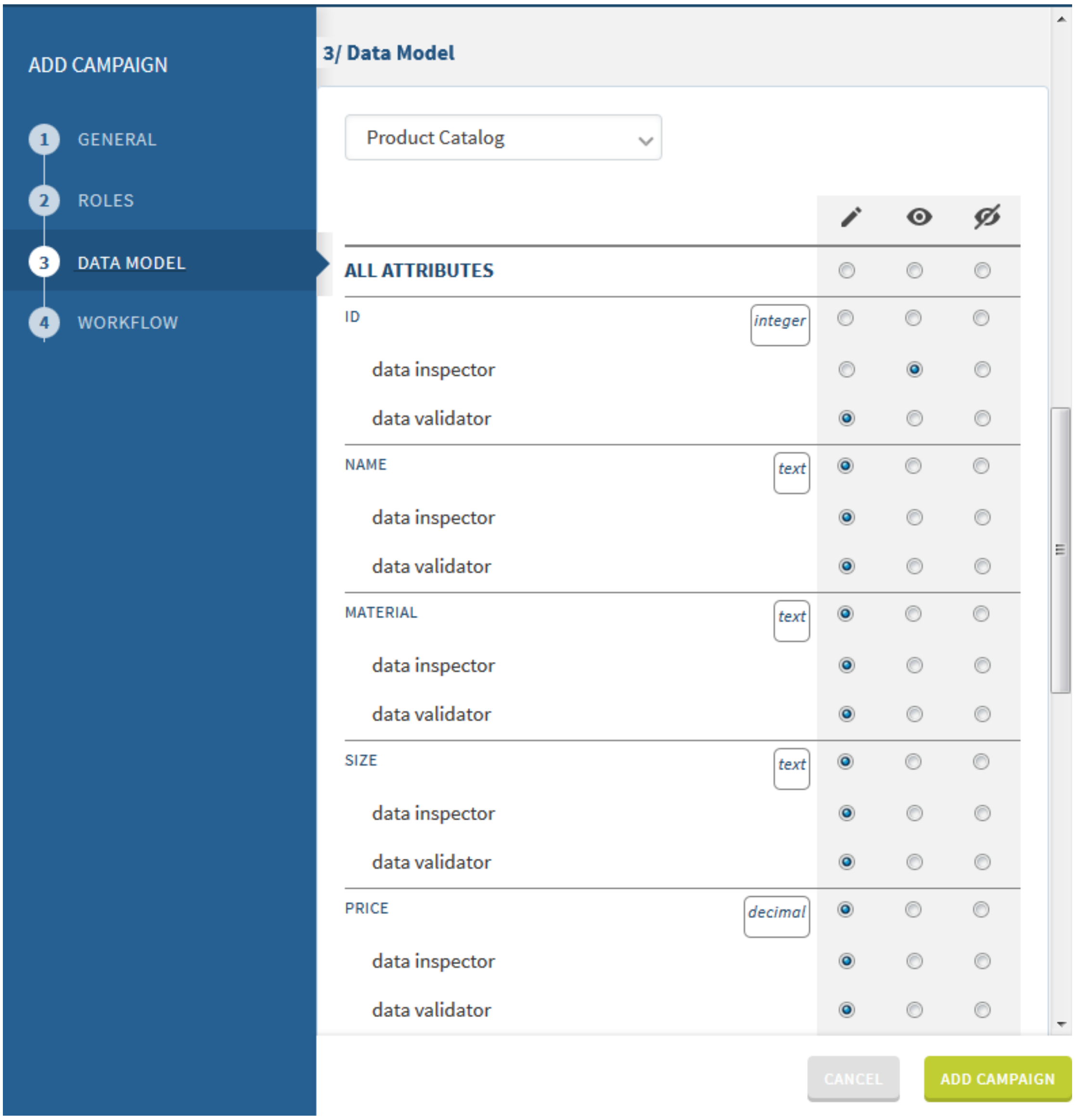
Setting a data model in the Resolution campaign

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 **Data Model** and select from the model list the data structure you want to use in the `Product Catalog` campaign.

   The **Data 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.

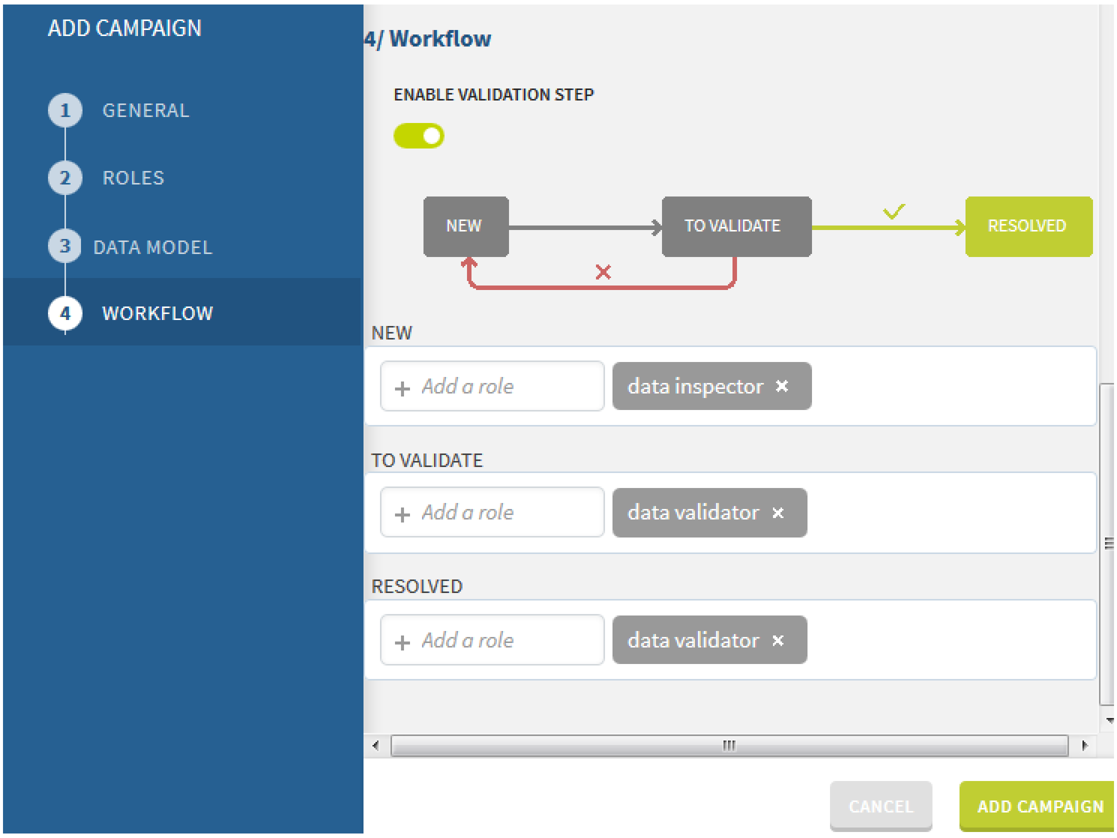
For example, here all participants can access and modify all the attributes in the data structure except the identifier for which the data stewards with the `data inspector` role have a read-only access.

Defining a workflow in the Resolution campaign

For the `Product Catalog` 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.

You can now:

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

For further information about writing tasks in campaigns, see the online publication on Talend Help Center (https://help.talend.com).

2. Assign tasks to data stewardships. For further information, see Assigning and reassigning tasks as a campaign owner on page 11.
3. Start tasks resolution.

   For further information, see Handling resolution tasks to curate product records on page 26.

Assigning and handling resolution tasks

Assigning and reassigning tasks as a campaign owner

Campaign owners can use the Job properties in Talend Studio to automatically assigned the tasks to a data steward. Otherwise, once logged to Talend Data Stewardship they can assign or reassign the tasks in a specific campaign to any data steward part of the campaign.

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

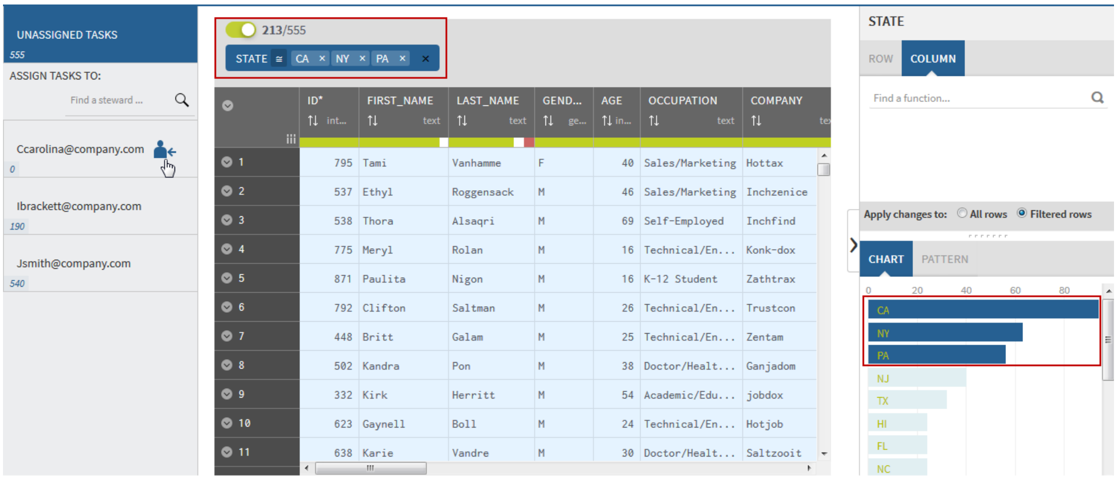
Assigning tasks

This procedure shows how to assign tasks manually.

For further information about assigning multiple tasks using a bulk operation, see the online publication on Talend Help Center (https://help.talend.com).

1. In the home page, select **CAMPAIGNS** and click the name of the campaign which holds the tasks you want to assign.
2. In the list of unassigned tasks, select a task or hold the Ctrl/SHIFT key and click to select multiple tasks and:
   - Drop the selected tasks to a data steward in the left panel, or
   - Point to the data steward to whom to assign the tasks and click next to his/her name.
   - Use the **ASSIGN TASKS TO** field in the top left corner to search for the data steward to whom to assign the tasks and click , or

   You can use the data quality bar on top of the columns or the charts displayed in the right-hand panel to filter data and list only the task you want to assign to specific data stewards.



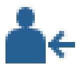
The selected tasks are moved to the task list of the data steward. The number of tasks which show under the participant name in the left panel is the number of tasks assigned to him/her in the current workflow state.

3. In the panel to the left, 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

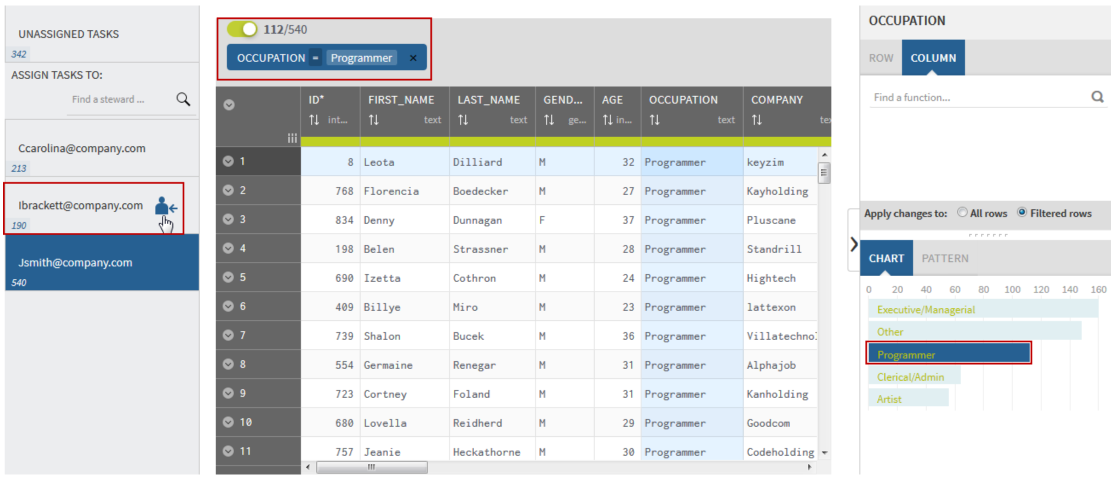
Reassigning tasks

Campaign owners can reassign tasks between the data stewards part of the campaign.

1. In the home page, select **CAMPAIGNS** and click the name of the campaign which holds the tasks you want to reassign.
2. In the left-hand panel, click a data steward to display the list of all the tasks assigned to him/her.
3. Select one or multiple tasks from the list and in the left-hand panel:

- Point to the data steward to whom to reassign the tasks and click  next to his/her name, or
- Use the **Assign Tasks To** field to search for the data steward to whom you want to reassign the tasks and click .

You can use the data quality bar on top of the columns or the charts displayed in the right-hand panel to filter data and list only the tasks you want to reassign.



The number of tasks which show under the user name in the left panel is the number of tasks assigned to the user within the campaign.

The selected tasks are moved from the initial data steward to the new data steward.

Assigning and delegating tasks as a data steward

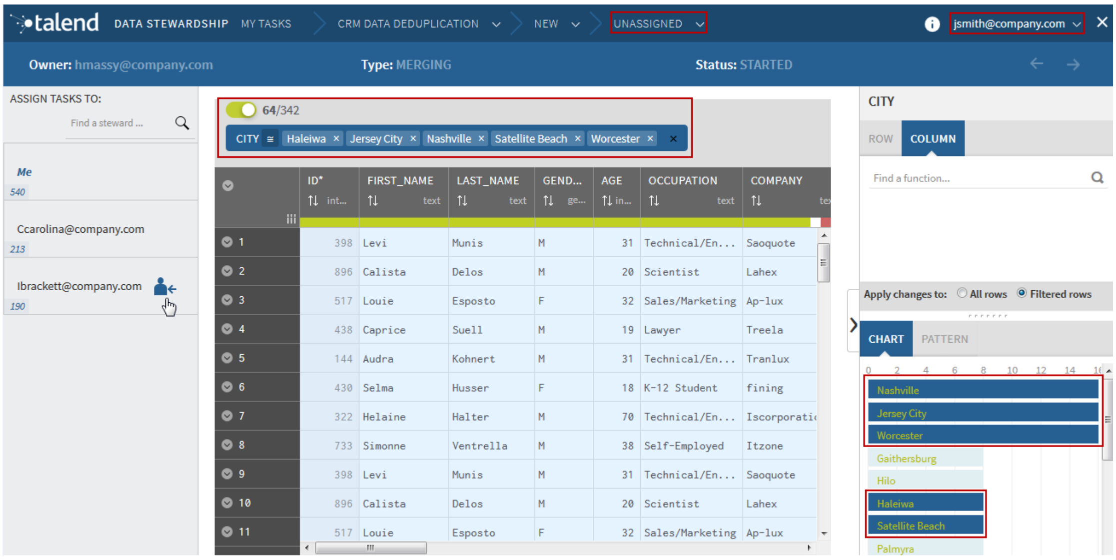
Data stewards can assign new tasks to themselves or other data stewards part of the same campaign. They can assign tasks to other data stewards only if tasks are currently unassigned or assigned to themselves.

Data stewards can assign or delegate tasks to other data stewards to meet certain needs, for example to improve the distribution of workload across business users or because of expertise on a specific task.

- A campaign owner has loaded the tasks in the campaign using a Talend Job.
- A campaign owner has created the campaign and granted you access to it.
- You have accessed Talend Data Stewardship as a data steward.

Assigning tasks

1. In the **TASKS** page, point to the campaign which holds the tasks you want to assign and click the icon that displays to the right of the campaign name.
   A list of all the unassigned tasks in the campaign is opened.



   You can click **Unassigned** on the menu bar and select **ASSIGNED TO ME** to open the list of the task already assigned to you in the campaign.
2. Use the data quality bar on top of the columns or the charts displayed in the right-hand panel to filter data and list only the tasks you want to assign, or select the tasks manually.

   For further information about filtering tasks with the quality bar or the charts, see the online publication on Talend Help Center (https://help.talend.com).
3. In the left-hand panel, either:

   - Point to the data steward to whom to assign the tasks and click next to his/her name, or

- Use the **Assign Tasks To** field to search for the data steward to whom you want to assign the tasks and click .

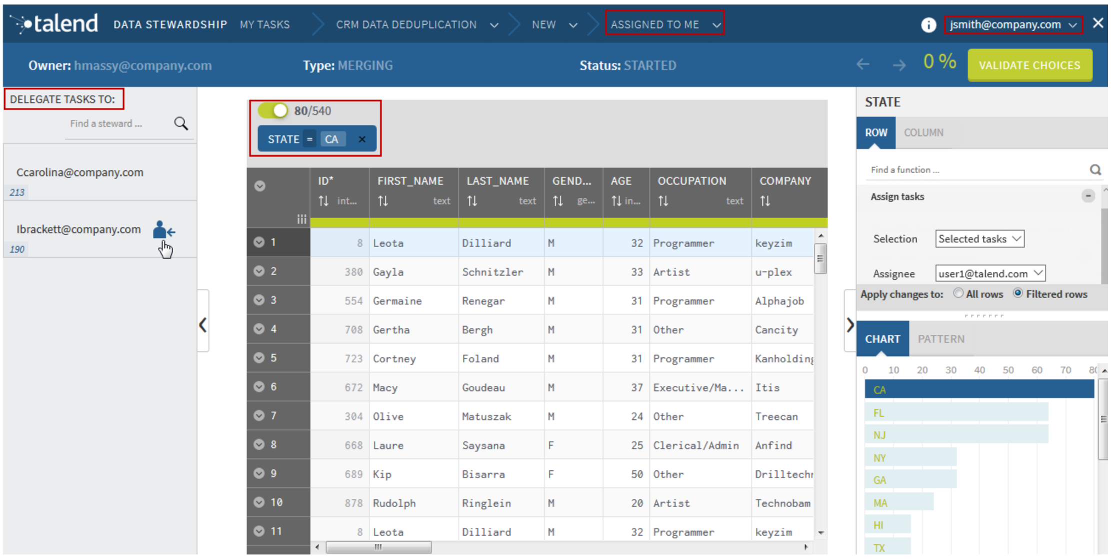
The selected tasks are moved to the dashboard of the data steward. The number of tasks which show under the participant name in the left panel is the number of tasks assigned to him/her within the campaign.

Delegating tasks

The reasons to delegate tasks include the need of certain expertise to handle the tasks, a risk of failing to do the tasks by the appointed time or fast approaching vacations.

Tasks have been assigned to the data stewards.

1. In the **TASKS** page, click the campaign which holds the tasks you want to delegate.



2. Use the chart in the right-hand panel to filter and select a group of the tasks assigned to you and in the left-hand panel, and do one of the following:
   - In the panel to the right, click **ROW** > **Assign tasks** and use the two lists to decide what tasks to delegate to what data steward.
   - Point to the data steward to whom to delegate the tasks and click next to his/her name.
   - Use the **Delegate Tasks To** field to search for the data steward to whom you want to assign the tasks and click next to his/her name.

Handling resolution tasks to curate product records

Resolution tasks aim to curate the records they contain. Data curation can consist in fixing one or several fields of the records. The outcome of a resolution task is the clean record produced by the data steward.

Curing data values and validating your modifications transition the task to the second state defined in the workflow. The workflow defined at the campaign creation determines which states are available to what data stewards. However, a task cannot be validated or even marked as ready as long as it contains at least one invalid value. And this guarantees that data which does not match your data model can go out of Talend Data Stewardship.

In this example, you need to approve product records and fix any anomalies to make sure the product catalog holds complete and accurate information. Once the records are cleansed, they can be uploaded to the Mater Data Management repository.

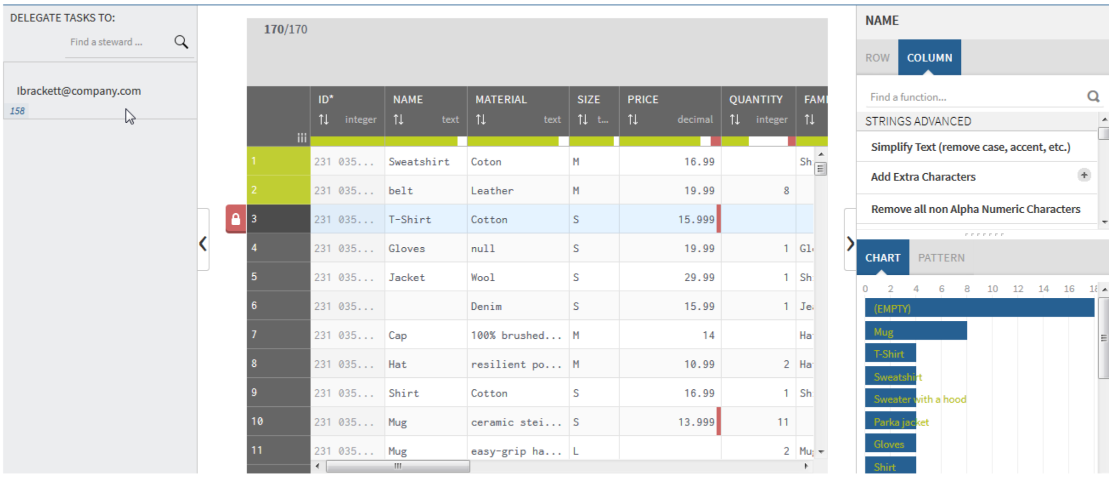
- A campaign owner has created the campaign and granted you access to it.

  For further information, see Adding a Resolution campaign to curate product records on page 15.

- A campaign owner has assigned you tasks in the campaign. Otherwise, you can assign the tasks to yourself.

  For further information, see the online publication about assigning tasks on Talend Help Center (https://help.talend.com).

1. In the **TASKS** page, click the campaign name, `Product Catalog` in this example, to open a list of the tasks assigned to you.



2. Click a color in the quality bar at the top of a column to filter the task list accordingly as the following:

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. Fix individual values or do a bulk fix for data in a specific column. Either:
   a) Double-click an invalid value in a column, enter a new value in the field and press **Enter**, or
   b) Use the functions in the right-hand panel to transform all the data in the selected column.

   For further information about the functions used to transform data in columns, see the online publication on Talend Help Center (https://help.talend.com)..

4. Repeat the above step to modify all invalid values and tagg curated tasks with green.

5. Click the  icon next to the data record you modified to mark the task with green background, i.e. ready to be validated.

6. If needed, click the curated task to display the lock icon and click the icon to put 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.
7. Click **VALIDATE CHOICES** in the top-right corner of the page to validate the modifications you have done on the records.
   In this example, the cleansed and validated records are moved from your list and transitioned to the **TO VALIDATE** step in the workflow. Data stewards who are granted the `DATA VALIDATOR` role can now access the tasks and approve them.
8. The data stewards with the `DATA VALIDATOR` role decide to accept or reject the choices done on the tasks.

Approved tasks are transitioned to the Resolved state in the workflow. Rejected tasks are transitioned back to the initial step in the workflow and marked as new.

Adding a Merging campaign to deduplicate records

A **Merging** campaign enables data stewards to merge several potential duplicate data records into one single master record. Source records can come from the same source (data deduplication) or different sources (data reconciliation).

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

One common use case of data deduplication is same customers appear as separate records in your CRM system. You would like here to match records in order to identify duplicates. A **MERGING** campaign enables you to decide what fields to use to determine a match and merge the records. Once data is deduplicated, a Talend Job can be used to reupload the cleansed data to CRM.

For a real world use case about data reconciliation, see the Getting started with Talend Data Stewardship publication on Talend Help Center (https://help.talend.com).

- 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 defined a data model for the campaign in Talend Data Stewardship.
- You have accessed Talend Data Stewardship as a campaign owner.

To add the campaign, you need to accomplish the below procedures:

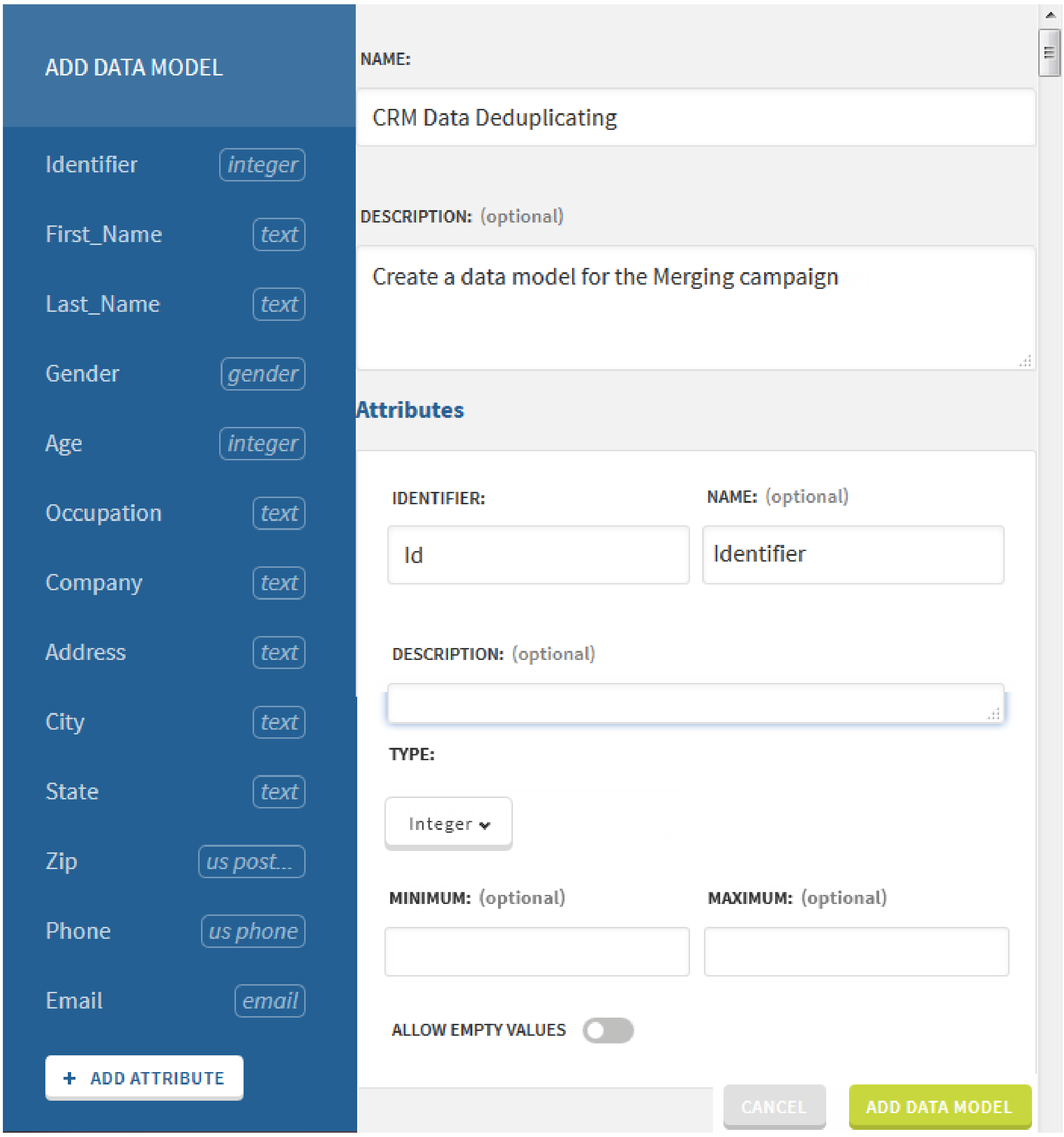
1. Defining a data model for the Merging campaign on page 28.
2. Defining the Merging campaign on page 30.
3. Defining roles in the Merging campaign on page 31.
4. Setting a data model in the Merging campaign on page 32.
5. Defining a workflow in the Merging campaign on page 34.

Defining a data model for the Merging campaign

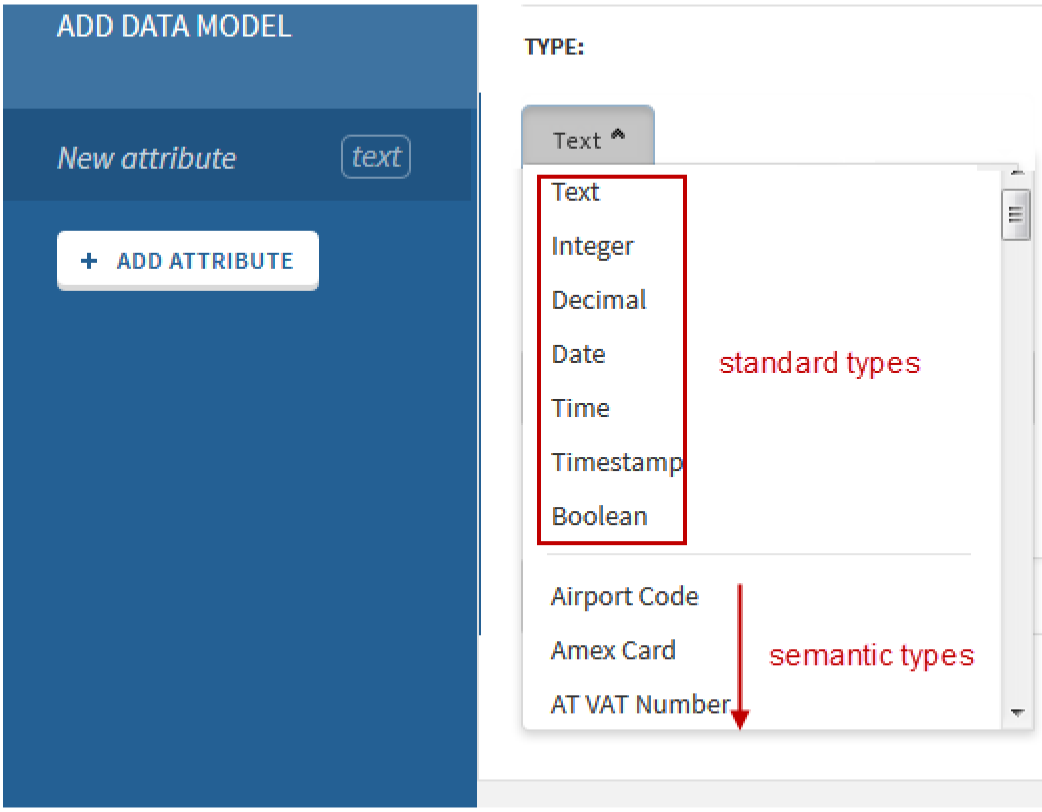
In this example, you create a data model to determine the structure of the data to be managed in the `CRM Data Deduplication` campaign which you create to allow data stewards to merge duplicate customer data stored in the enterprise CRM.

Talend Data Stewardship has data model awareness which makes possible the syntactic and semantic validation of data. You can define the attributes in the data model and select their types out of a predefined standard or semantic types.

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.



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.

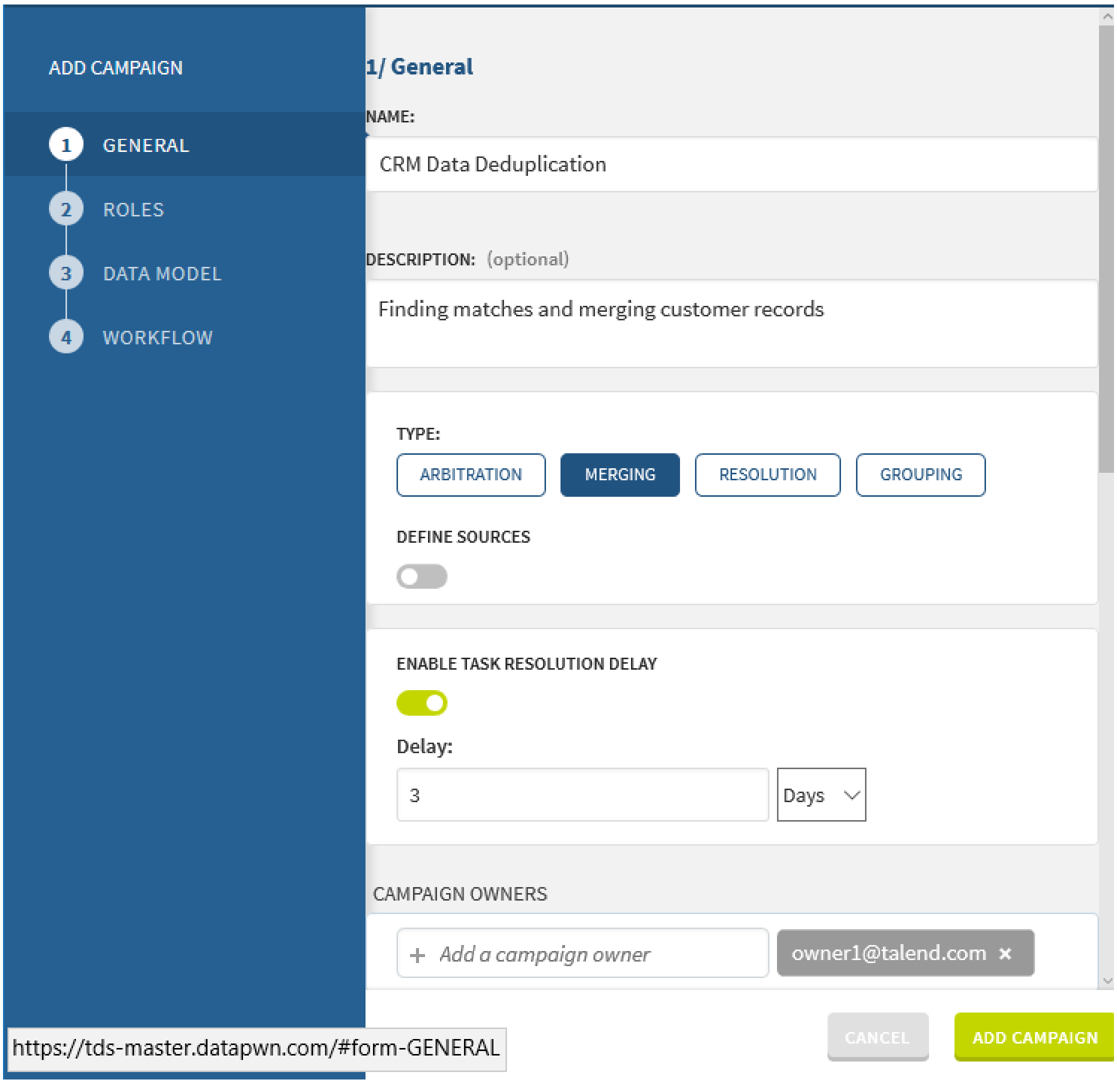
4. If needed, click the switch next to **ALLOW EMPTY VALUES** to disable the load of empty fields to Talend Data Stewardship. This option is enabled by default.
5. Click **ADD ATTRIBUTE** in the left panel and repeat the above steps to create all the columns you need in the data model.
   The columns defined for the `CRM Data Deduplication` campaign include information about the customers and the company in which they work as shown in the capture.

Defining the Merging campaign

In this scenario, you define the `CRM Data Deduplication` campaign to allow data stewards to address potential duplicates in the contact records in the CRM system of an enterprise.

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. Leave the **DEFINE SOURCES** option inactivated as duplicate data in this example comes from one source.
   When you define source names in the campaign, make sure that they do not contain dots and that they do not start with a dollar signs.
5. 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.
6. 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.

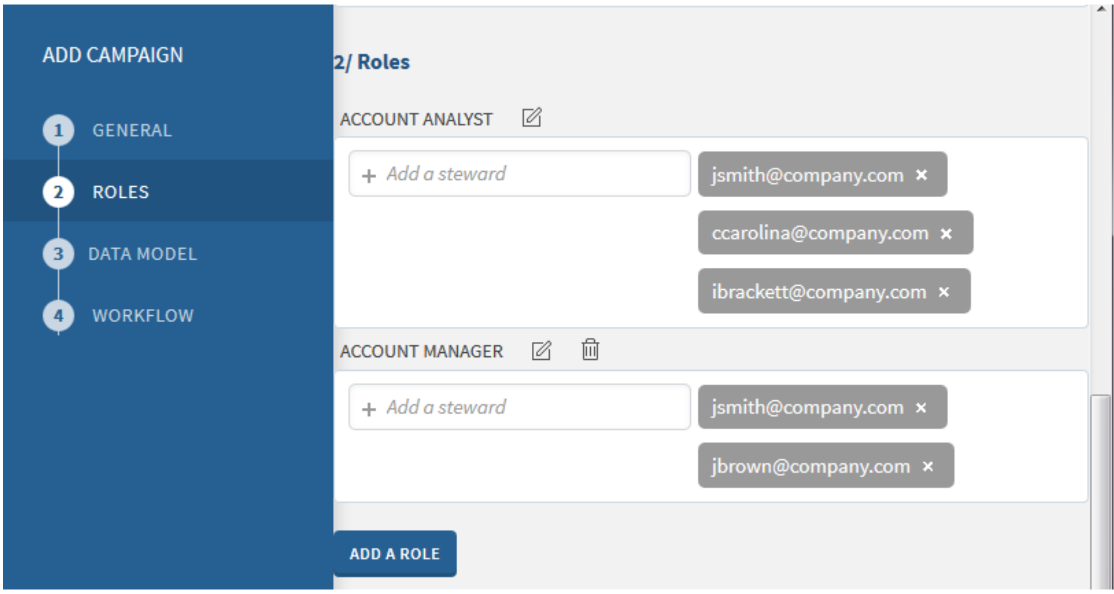
Defining roles in the Merging campaign

You need now to define roles in the campaign to decide the actions to be taken on data, which data stewards are part of the campaign and which data stewards can access which tasks and in what transition of the campaign workflow.

You can create and start a campaign without defining any roles and edit the campaign later to add the roles of your choice.

After defining the metadata of the campaign, you need to group data stewards into roles and assign them access rights.

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



3. 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.
4. Repeat the above steps to add a second role and name it `ACCOUNT MANAGER`.
5. Select from the **Add a steward** list the user to validate the choices of the user who is assigned the first role.

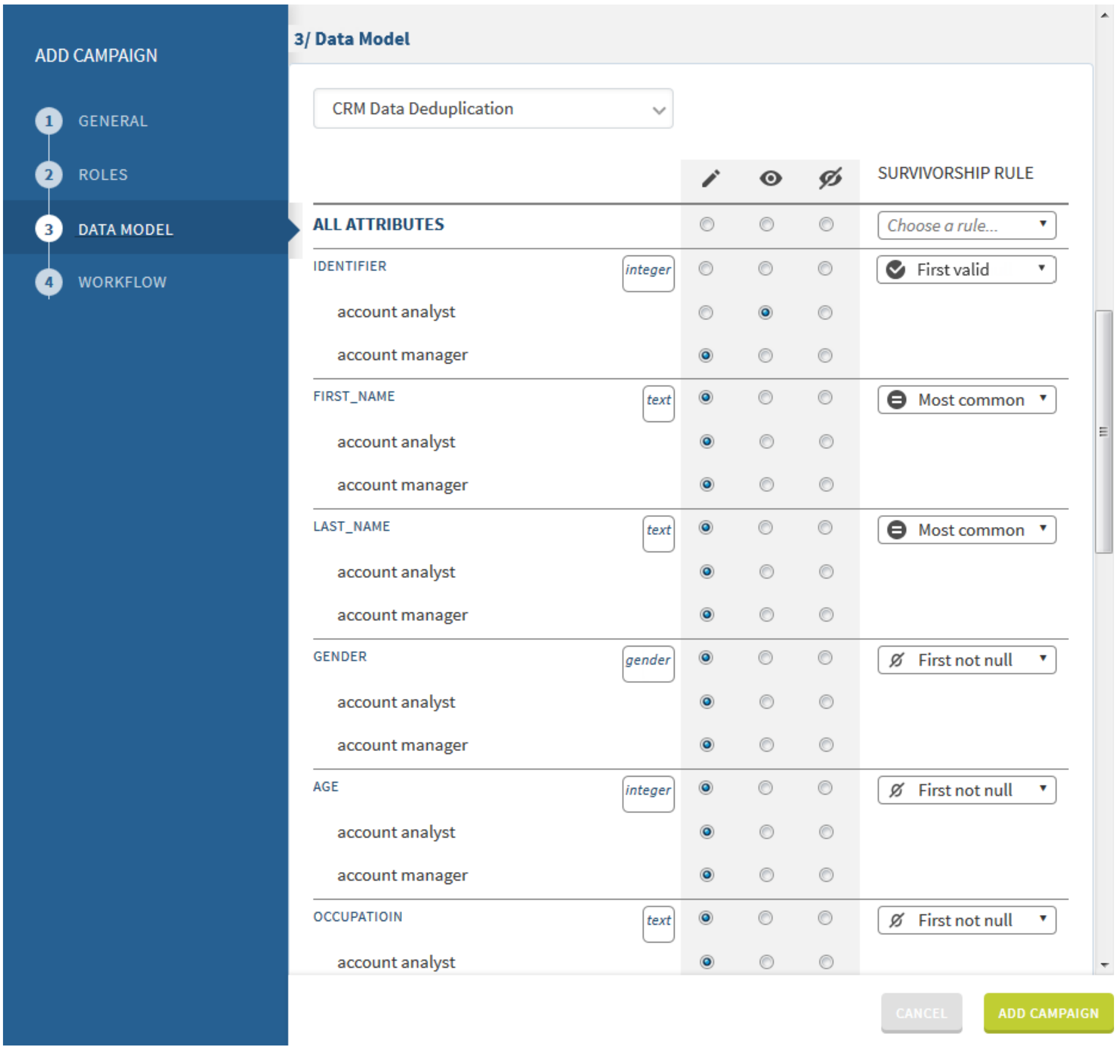
Setting a data model in the Merging campaign

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 **Data Model** and select from the model list the data structure you want to use in the `CRM Data Deduplication` campaign.

   The **Data 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 the `CRM Data Deduplication` campaign, you grant a read-only access to the identifier attribute for the data stewards who are assigned the `account analyst` role.

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.

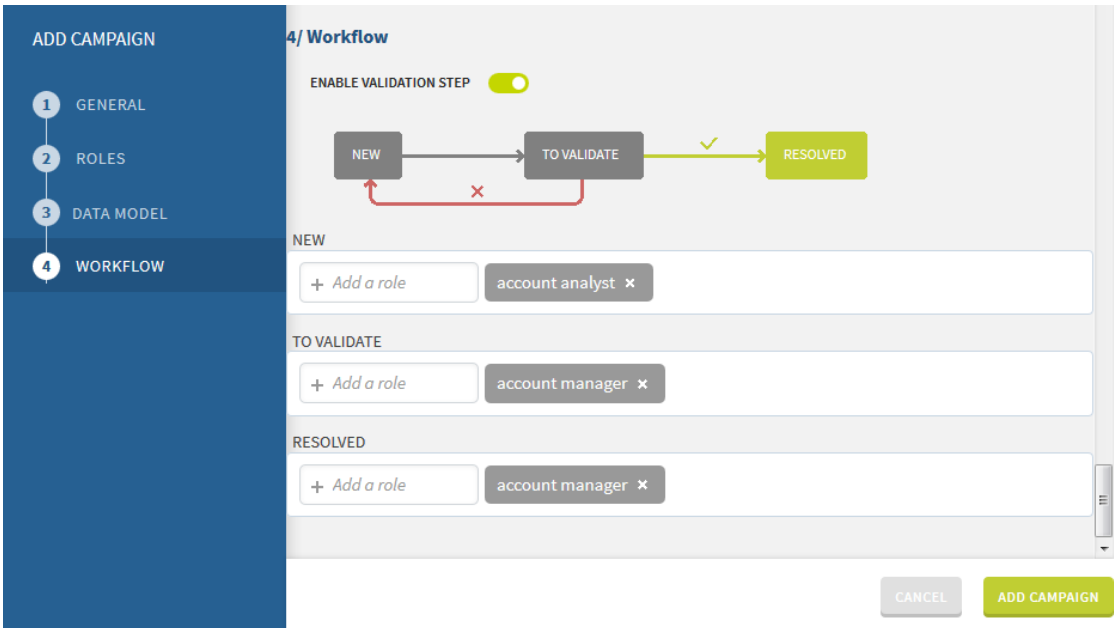
You can select one rule for all the attributes by selecting it from the list in the top-right corner of the form. If a given algorithm cannot be applied, the rule falls back to **First not null**. For example, if you do not set a trust score and you select **Most trusted** during the campaign definition, **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 Merging campaign

For the `CRM Data Deduplication` 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.

You can now:

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

   For further information, see the online publication about writing tasks in **Merging** campaigns on Talend Help Center (https://help.talend.com).
2. Assign tasks to data stewardships. For further information, see Assigning and reassigning tasks as a campaign owner on page 11.
3. Start tasks resolution.

   For further information, see Handling merging tasks to deduplicate records on page 38.

Assigning and handling merging tasks

Assigning and reassigning tasks as a campaign owner

Campaign owners can use the Job properties in Talend Studio to automatically assigned the tasks to a data steward. Otherwise, once logged to Talend Data Stewardship they can assign or reassign the tasks in a specific campaign to any data steward part of the campaign.

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

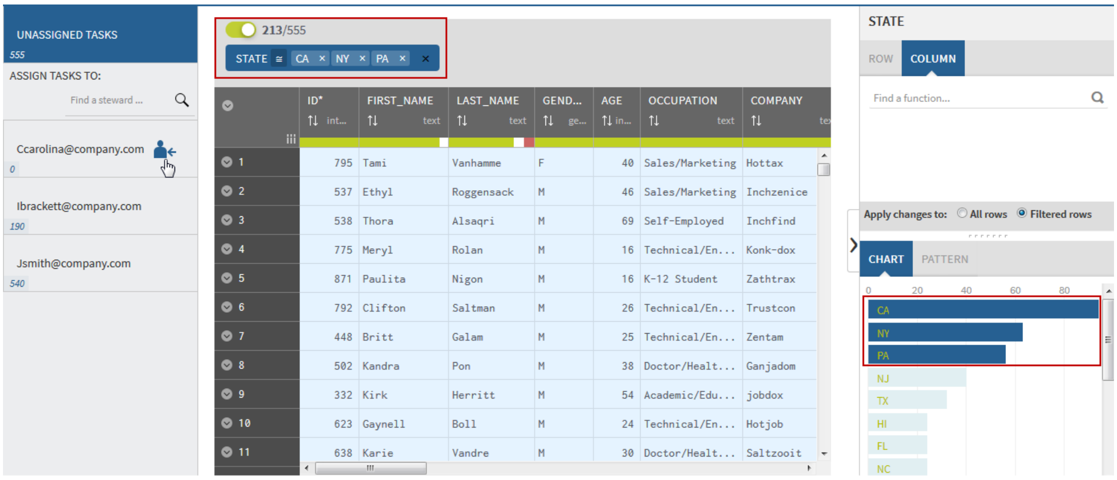
Assigning tasks

This procedure shows how to assign tasks manually.

For further information about assigning multiple tasks using a bulk operation, see the online publication on Talend Help Center (https://help.talend.com).

1. In the home page, select **CAMPAIGNS** and click the name of the campaign which holds the tasks you want to assign.
2. In the list of unassigned tasks, select a task or hold the Ctrl/SHIFT key and click to select multiple tasks and:
   - Drop the selected tasks to a data steward in the left panel, or
   - Point to the data steward to whom to assign the tasks and click next to his/her name.
   - Use the **ASSIGN TASKS TO** field in the top left corner to search for the data steward to whom to assign the tasks and click , or

   You can use the data quality bar on top of the columns or the charts displayed in the right-hand panel to filter data and list only the task you want to assign to specific data stewards.



The selected tasks are moved to the task 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 current workflow state.

3. In the panel to the left, 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

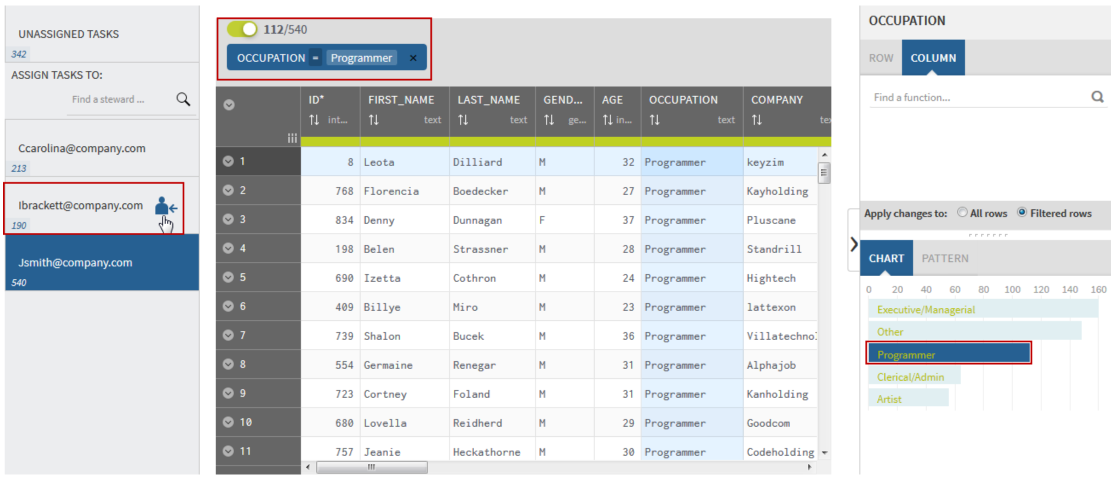
Reassigning tasks

Campaign owners can reassign tasks between the data stewards part of the campaign.

1. In the home page, select **CAMPAIGNS** and click the name of the campaign which holds the tasks you want to reassign.
2. In the left-hand panel, click a data steward to display the list of all the tasks assigned to him/her.
3. Select one or multiple tasks from the list and in the left-hand panel:

- Point to the data steward to whom to reassign the tasks and click  next to his/her name, or
- Use the **Assign Tasks To** field to search for the data steward to whom you want to reassign the tasks and click .

You can use the data quality bar on top of the columns or the charts displayed in the right-hand panel to filter data and list only the tasks you want to reassign.



The number of tasks which show under the user name in the left panel is the number of tasks assigned to the user within the campaign.

The selected tasks are moved from the initial data steward to the new data steward.

Assigning and delegating tasks as a data steward

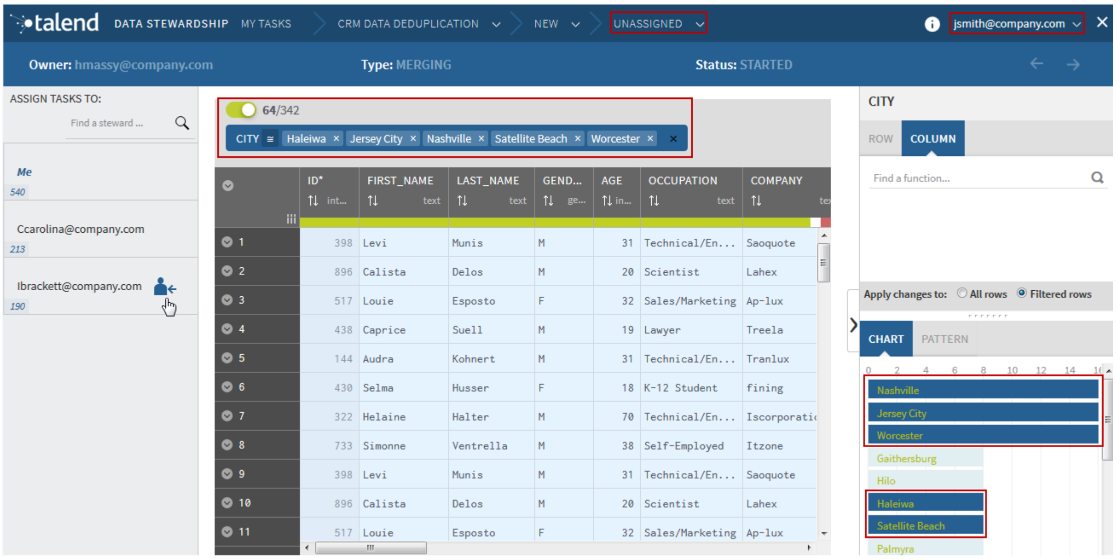
Data stewards can assign new tasks to themselves or other data stewards part of the same campaign. They can assign tasks to other data stewards only if tasks are currently unassigned or assigned to themselves.

Data stewards can assign or delegate tasks to other data stewards to meet certain needs, for example to improve the distribution of workload across business users or because of expertise on a specific task.

- A campaign owner has loaded the tasks in the campaign using a Talend Job.
- A campaign owner has created the campaign and granted you access to it.
- You have accessed Talend Data Stewardship as a data steward.

Assigning tasks

1. In the **TASKS** page, point to the campaign which holds the tasks you want to assign and click the icon that displays to the right of the campaign name.
   A list of all the unassigned tasks in the campaign is opened.



   You can click **Unassigned** on the menu bar and select **ASSIGNED TO ME** to open the list of the task already assigned to you in the campaign.

2. Use the data quality bar on top of the columns or the charts displayed in the right-hand panel to filter data and list only the tasks you want to assign, or select the tasks manually.

   For further information about filtering tasks with the quality bar or the charts, see the online publication on Talend Help Center (https://help.talend.com).

3. In the left-hand panel, either:

   - Point to the data steward to whom to assign the tasks and click  next to his/her name, or

- Use the **Assign Tasks To** field to search for the data steward to whom you want to assign the tasks and click .

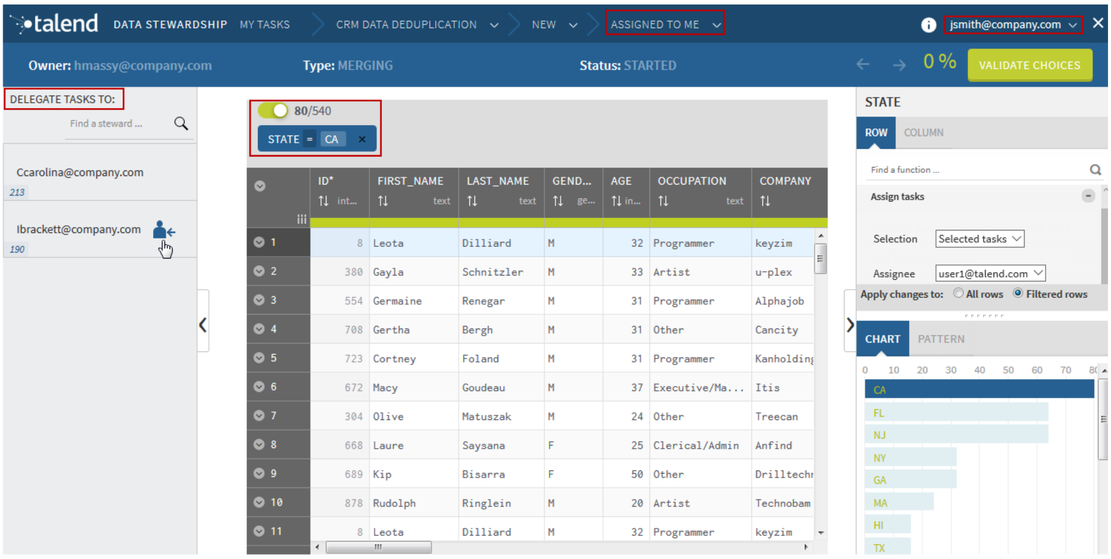
The selected tasks are moved to the dashboard of the data steward. The number of tasks which show under the participant name in the left panel is the number of tasks assigned to him/her within the campaign.

Delegating tasks

The reasons to delegate tasks include the need of certain expertise to handle the tasks, a risk of failing to do the tasks by the appointed time or fast approaching vacations.

Tasks have been assigned to the data stewards.

1. In the **TASKS** page, click the campaign which holds the tasks you want to delegate.



2. Use the chart in the right-hand panel to filter and select a group of the tasks assigned to you and in the left-hand panel, and do one of the following:

   - In the panel to the right, click **ROW** > **Assign tasks** and use the two lists to decide what tasks to delegate to what data steward.
   - Point to the data steward to whom to delegate the tasks and click next to his/her name.
   - Use the **Delegate Tasks To** field to search for the data steward to whom you want to assign the tasks and click next to his/her name.

Handling merging tasks to deduplicate records

Merging tasks aim to merge several potential duplicate records into one single record: master record. Potential duplicate records can come from the same source (data deduplication) or from different sources (data reconciliation).

In a **Merging** campaign, you can only modify values in the master fields, values in the source fields can not be modified.

Merging data values and validating your modifications transition the task to the second state defined in the workflow. The workflow defined at the campaign creation determines which states are available to what data stewards. However, a task cannot be validated or even marked as ready as long as it contains at least one invalid value. And this guarantees that data which does not match the data model can not go out of Talend Data Stewardship.

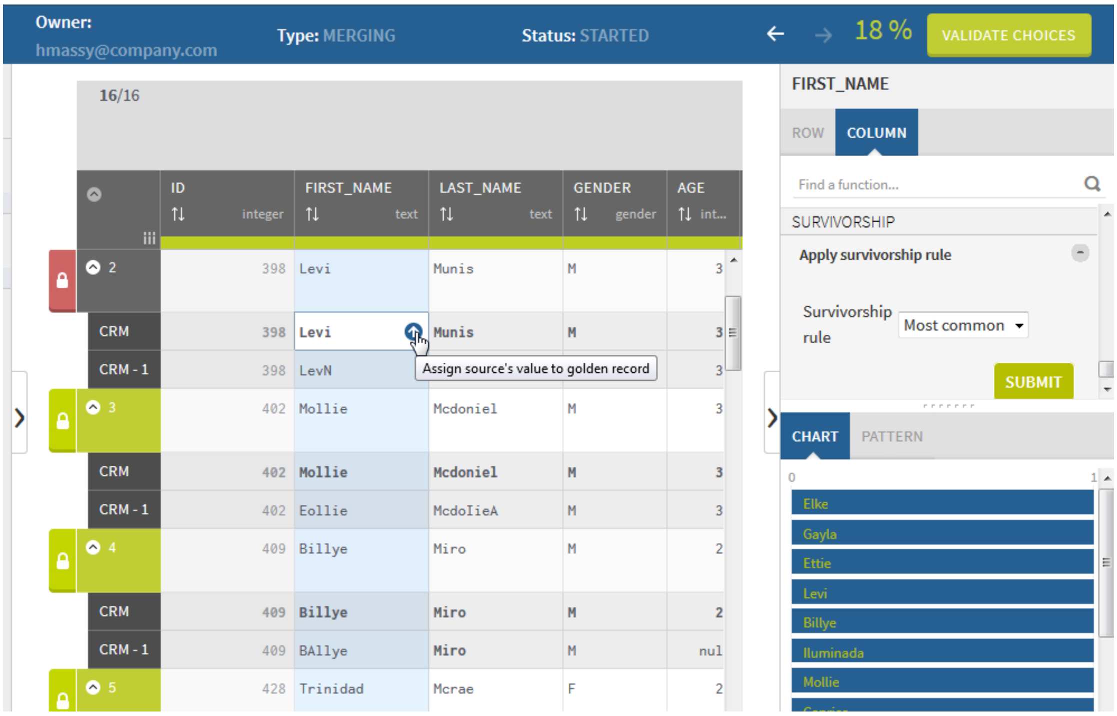
In this example, customer duplicate records come from the same source (enterprise CRM). Talend Data Stewardship determines initially which attributes of matched records to use to create the master record according to the survivorship rules defined when creating the campaign. However, you may need to 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 Adding a Merging campaign to deduplicate records on page 28.
- A campaign owner has assigned you tasks in the campaign. Otherwise, you can assign the tasks to yourself.

  For further information, see the online publication about assigning tasks on Talend Help Center (https://help.talend.com).

1. In the **TASKS** page, click the campaign name, `CRM Data Deduplication` in this example, to open a list of the tasks assigned to you.



2. Use the quality bar on top of each of the columns to filter the data on which you want to work in the **CHARTS** or **PATTERN** views in the right-hand panel.
3. Click the down arrow on the top-left corner to expand all tasks in the list, 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 coming from different sources.

  Icons are grayed out when rules are not applicable on the selected attribute. In this example, the icon for the most trusted attribute is not functional since customer data comes from one single source: CRM.
- Set a survivorship rule manually for one attribute of multiple records.
  1. Click a column heading, `First_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 customer records.
  3. Click **Submit** to select the most common name values and add them to the master records of 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. Repeat the above step to merge records and create master records for all the tasks assigned to you. If a given column has some values which need to be fixed, you can bulk transform them by using the functions listed in the right panel.

   For further information about the functions used to transform data in columns, see the online publication on Talend Help Center (https://help.talend.com)..
6. Click the icon next to the data record you modified to mark the task as ready to be validated. When the lock icon has a red background color, you must first correct the invalid value in the task before being able to mark it as ready to be validated.

   The record is marked with green background and the lock icon is automatically moved to the next record. 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.
7. Click **VALIDATE CHOICES** in the top-right corner of the page to validate the modifications you have done on the records.
   Master records are created and the records which are validated are moved from the list and transitioned to the **TO VALIDATE** step in the workflow where they need to be approved by another data steward. In this example, they are moved to the list of the data steward who is granted the `ACCOUNT MANGAGER` role.
8. The data stewards with the `ACCOUNT MANAGER` role, access the tasks to be validated and decide to accept or reject the choices done on the tasks.

Approved tasks are transitioned to the Resolved state in the workflow. Rejected tasks are transitioned back to the initial step in the workflow and marked as new.

Adding a Grouping campaign to identify duplicate pairs

A **Grouping** campaign defines a list of possible arbitration choices for pairs or groups of records. The outcome of a grouping task is the choice made by data stewards on the group of records.

A typical use case for this campaign is to label suspect duplicate pairs in the context of matching very high volume of data using machine learning on Spark. Another use case for the **Grouping** campaign is to identify the groups of potential duplicates before sending them to a **Merging** campaign where data stewards can merge duplicates into master records.

The use case described here uses a **Grouping** campaign to identify duplicates in a sample data extracted from a product catalog. Once campaign owners create the campaign, data stewards need to look at the sample data and say whether pairs of record are duplicate or not.

- 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 defined a data model for the campaign in Talend Data Stewardship.
- You have accessed Talend Data Stewardship as a campaign owner.

To add the campaign, you need to accomplish the below procedures:

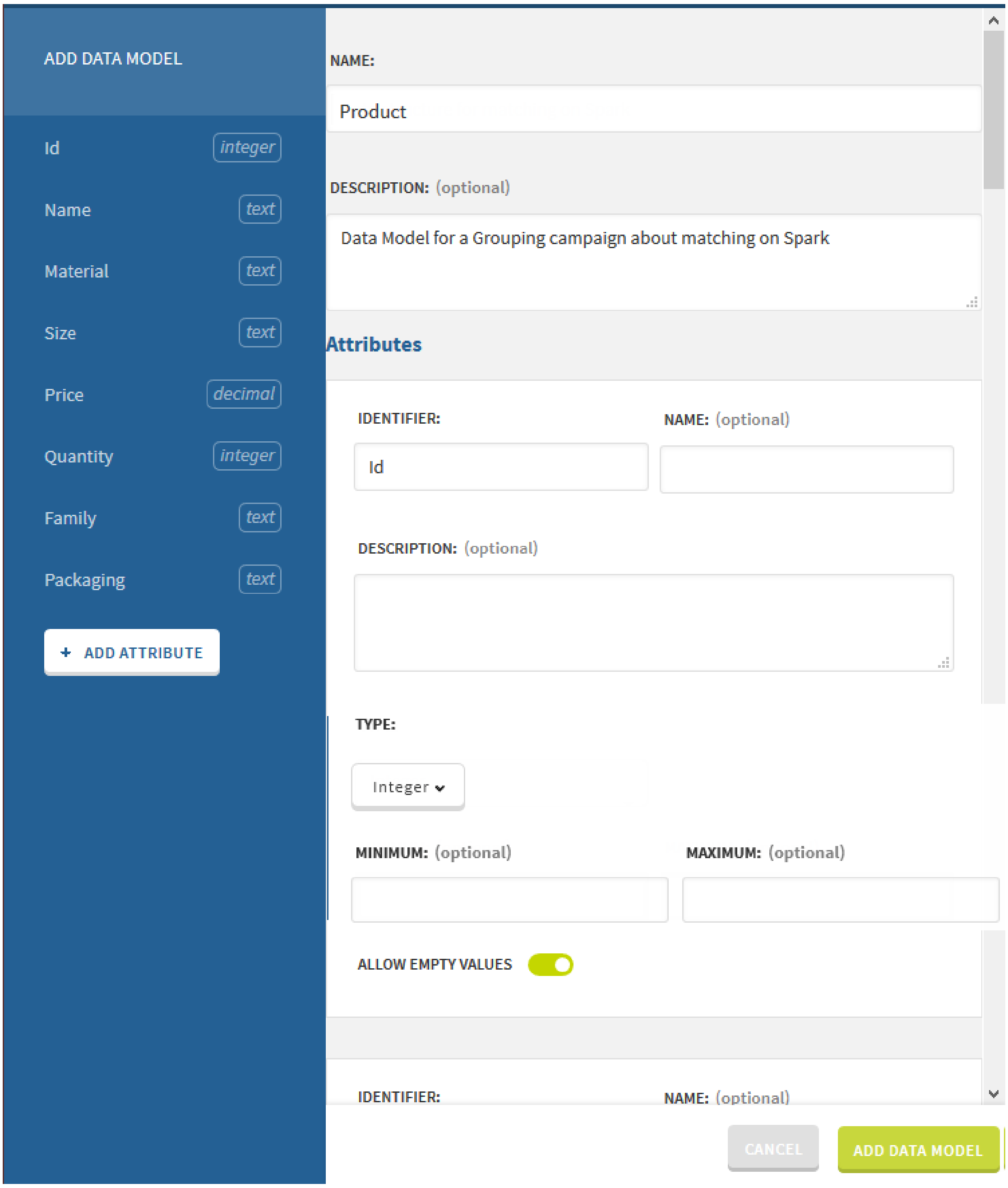
1. Defining a data model for the grouping campaign on page 41.
2. Defining the Grouping campaign on page 43.
3. Defining roles in the Grouping campaign on page 45.
4. Setting the data model in the Grouping campaign on page 45.
5. Defining a workflow in the Grouping campaign on page 47.

Defining a data model for the grouping campaign

In this example, you create a data model to determine the structure of the data to be managed in the `Product deduplication` campaign you create to label near duplicates in a data sample extracted by a Talend Job.

Talend Data Stewardship has data model awareness which makes possible the syntactic and semantic validation of data. You can define the attributes in the data model and select their types out of a predefined standard or semantic types.

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.

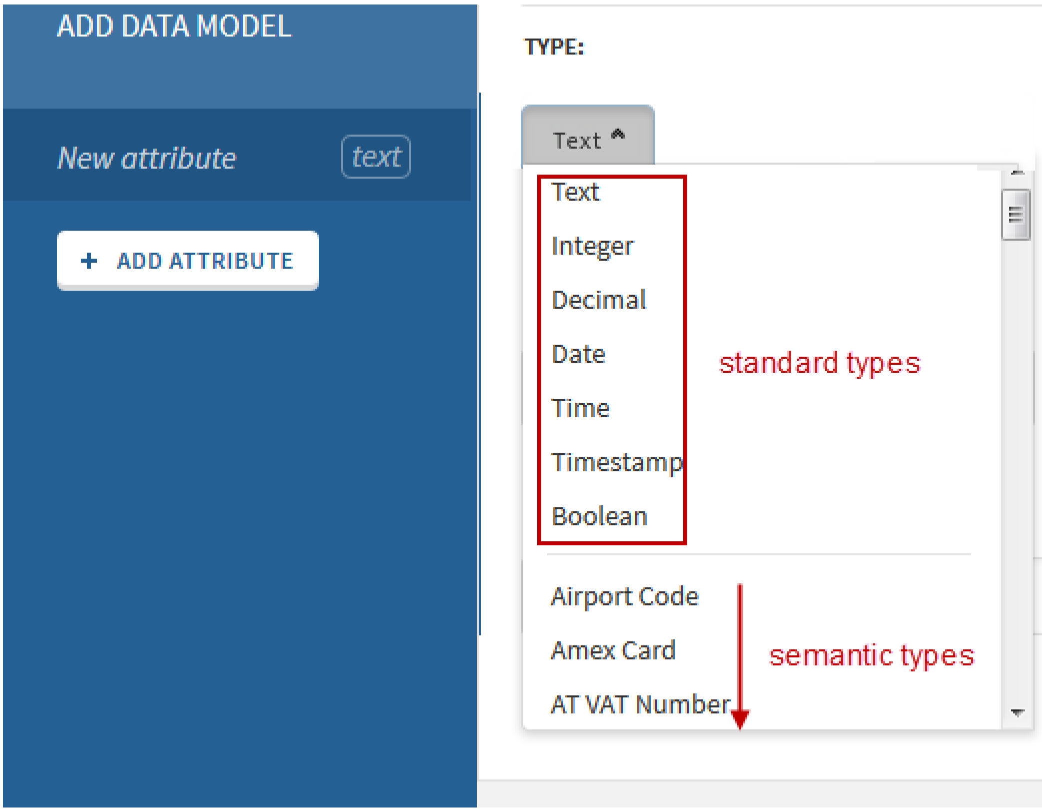


3. In the **Attributes** section, define the columns you want to have in the data model as the following:

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.

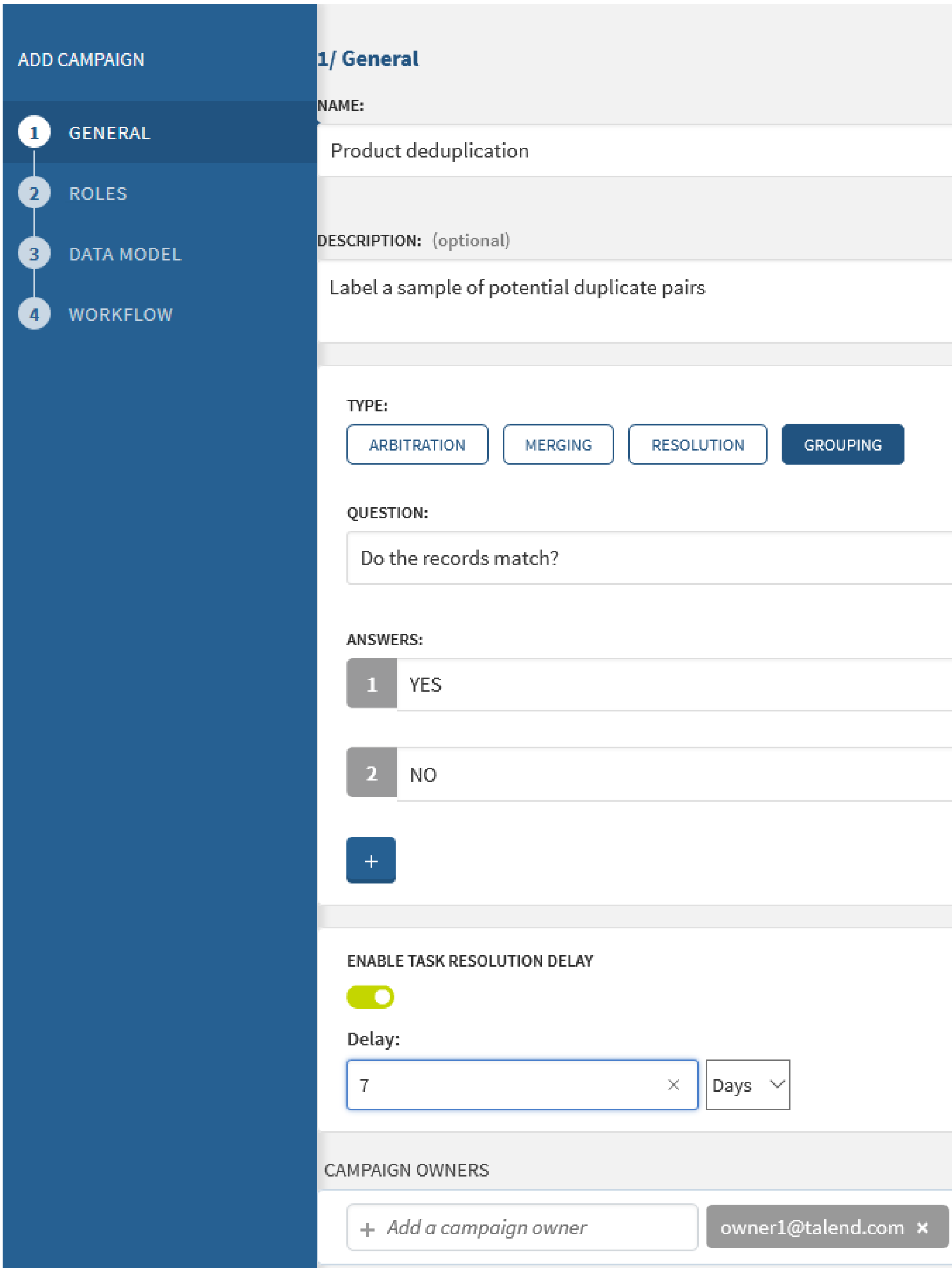
4. If needed, click the switch next to **ALLOW EMPTY VALUES** to disable the load of empty fields to Talend Data Stewardship. This option is enabled by default.
5. Click **ADD ATTRIBUTE** in the left panel and repeat the above steps to create all the columns you need in the data model.
   The columns defined for the `Product deduplication` campaign used in this example hold descriptive information about product items including their size, price and packaging as shown in the capture.

Defining the Grouping campaign

In this scenario, you define the `Product deduplication` campaign to allow data stewards to label a sample of suspect records in the context of matching very high volume of data using machine learning on Spark.

Start by defining the campaign metadata.

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



4. In the **QUESTION** field, enter the question to use in the campaign, for example: `Do the records match?`
5. Use the  button to add one more choice, and enter the answers in the **ANSWERS** fields as shown in the above capture.
6. 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 **Grouping** campaign is the most recent date of all the records.

7. 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.

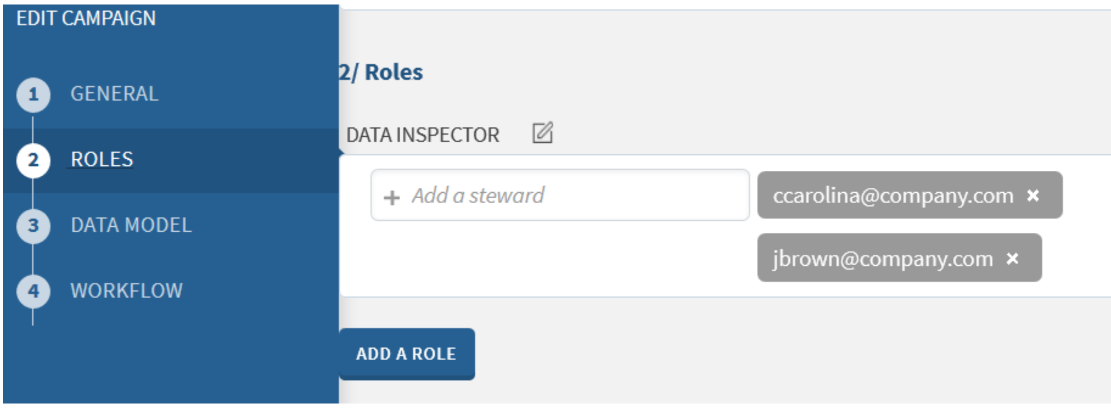
Defining roles in the Grouping campaign

You need now to define roles in the campaign to decide the actions to be taken on data, which data stewards are part of the campaign and which data stewards can access which tasks and in what transition of the campaign workflow.

You can create and start a campaign without defining any roles and edit the campaign later to add the roles of your choice.

After defining the metadata of the campaign, you need to group data stewards into roles and assign them access rights.

1. In the home page, click **ROLES** > .
2. Name the role `DATA INSPECTOR` and click next to the field to save it.



3. 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.

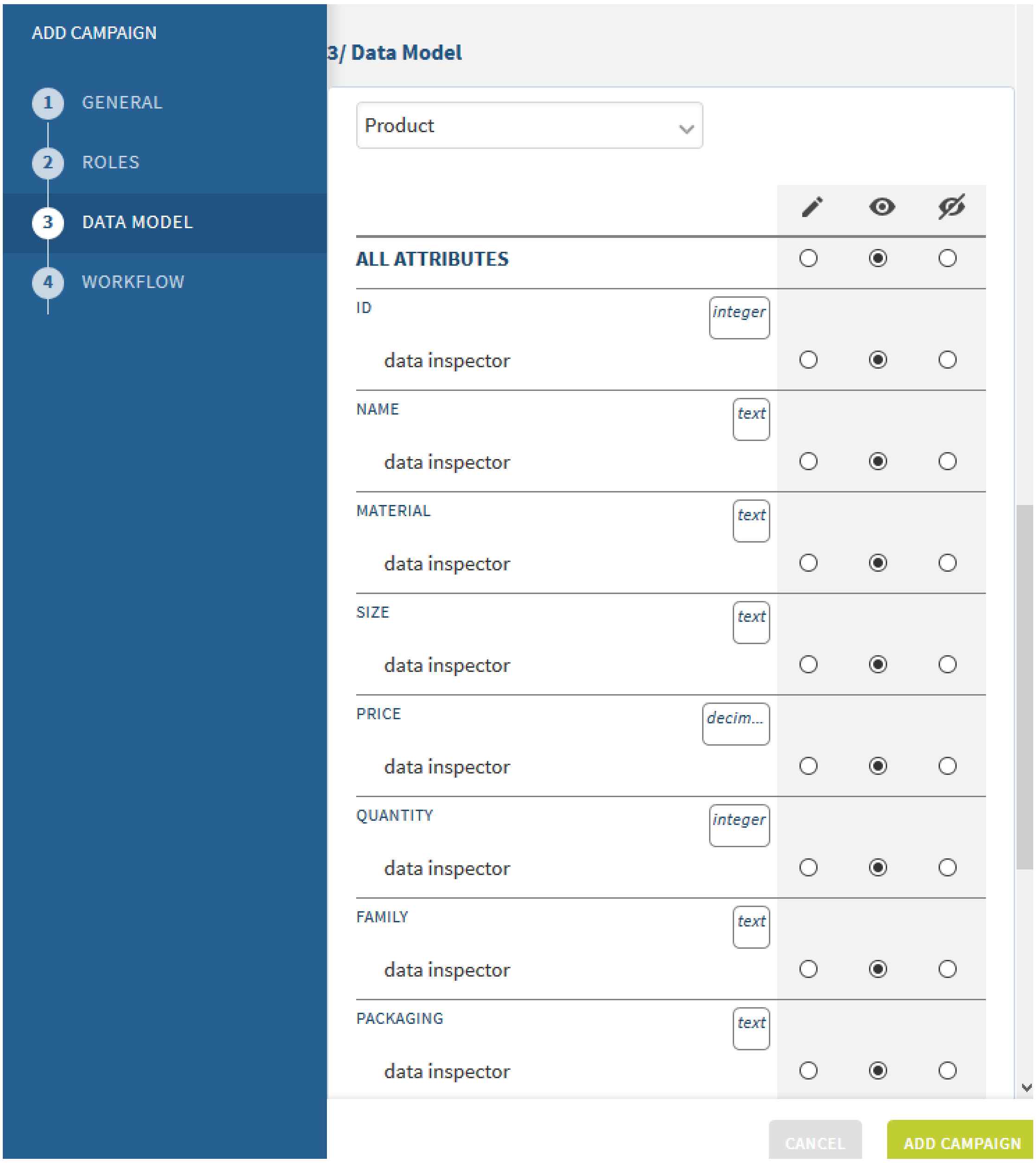
Setting the data model in the Grouping campaign

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 **Data Model** and select from the model list the data structure you want to use in the `Product deduplication` campaign.

   The **Data 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.

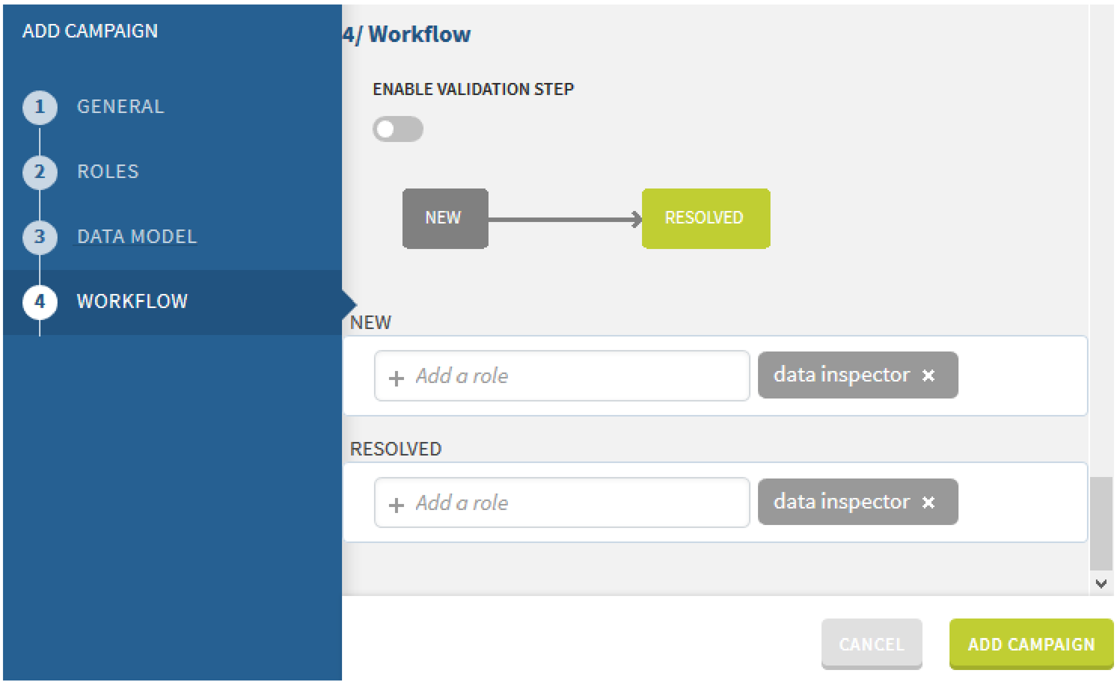
In this example you have only the `data inspector` role. You want to grant data stewards a read-only access to all attributes as data stewards only need to analyze pairs of records and decide if they are near duplicates.

Defining a workflow in the Grouping campaign

For the `Product deduplication` campaign, you want to use the by-default workflow which has one initial state and one final state. The newly created tasks are assigned to specific data stewards, handled and then marked 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 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.



2. Do the same in the **RESOLVED** field and select the roles which you want to grant access to the resolved tasks.

   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).
3. Click **ADD CAMPAIGN** to create the campaign and save it on the server.

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

   For further information about the tMatchPairing component, see the online publications at https://help.talend.com/.
2. Assign tasks to data stewardships. For further information, see Assigning and reassigning tasks as a campaign owner on page 11.

3. Start tasks resolution.

For further information, see Handling grouping tasks to decide on relationship among pairs of records on page 51.

Assigning and handling grouping tasks

Assigning and reassigning tasks as a campaign owner

Campaign owners can use the Job properties in Talend Studio to automatically assigned the tasks to a data steward. Otherwise, once logged to Talend Data Stewardship they can assign or reassign the tasks in a specific campaign to any data steward part of the campaign.

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

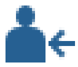
Assigning tasks

This procedure shows how to assign tasks manually.

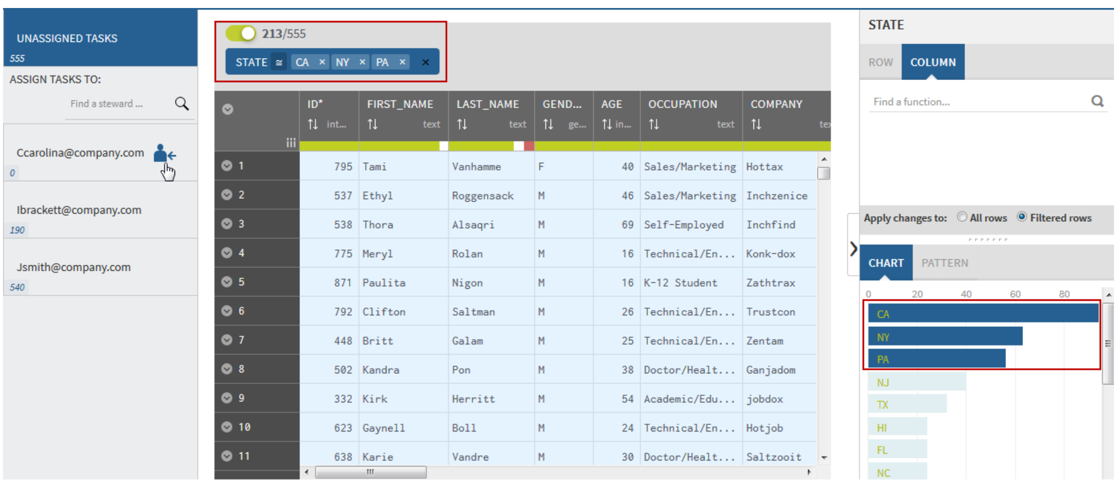
For further information about assigning multiple tasks using a bulk operation, see the online publication on Talend Help Center (https://help.talend.com).

1. In the home page, select **CAMPAIGNS** and click the name of the campaign which holds the tasks you want to assign.

2. In the list of unassigned tasks, select a task or hold the Ctrl/SHIFT key and click to select multiple tasks and:

- Drop the selected tasks to a data steward in the left panel, or
- Point to the data steward to whom to assign the tasks and click  next to his/her name.
- Use the **ASSIGN TASKS TO** field in the top left corner to search for the data steward to whom to assign the tasks and click , or

You can use the data quality bar on top of the columns or the charts displayed in the right-hand panel to filter data and list only the task you want to assign to specific data stewards.



The selected tasks are moved to the task 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 current workflow state.

3. In the panel to the left, 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

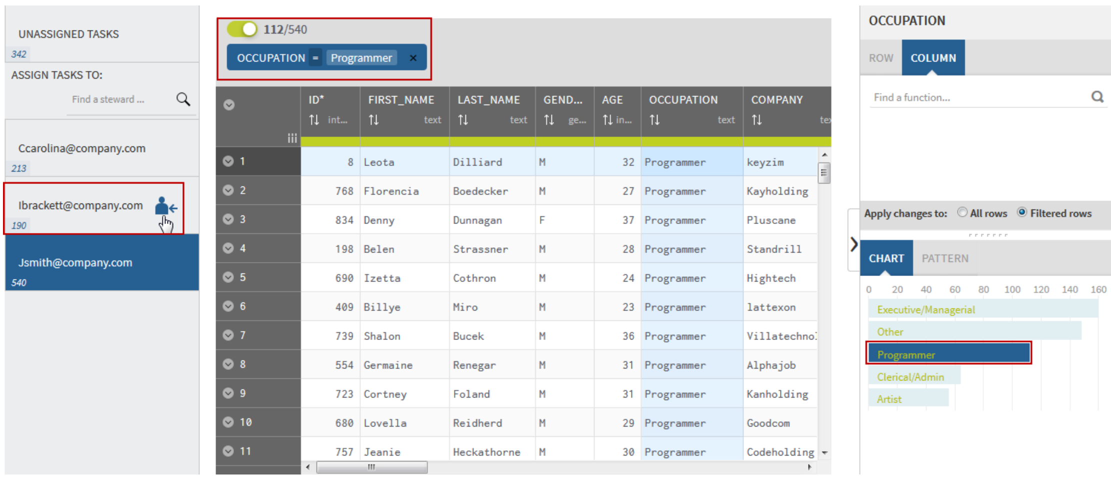
Reassigning tasks

Campaign owners can reassign tasks between the data stewards part of the campaign.

1. In the home page, select **CAMPAIGNS** and click the name of the campaign which holds the tasks you want to reassign.
2. In the left-hand panel, click a data steward to display the list of all the tasks assigned to him/her.
3. Select one or multiple tasks from the list and in the left-hand panel:

   - Point to the data steward to whom to reassign the tasks and click next to his/her name, or
   - Use the **Assign Tasks To** field to search for the data steward to whom you want to reassign the tasks and click .

   You can use the data quality bar on top of the columns or the charts displayed in the right-hand panel to filter data and list only the tasks you want to reassign.



   The number of tasks which show under the user name in the left panel is the number of tasks assigned to the user within the campaign.

The selected tasks are moved from the initial data steward to the new data steward.

Assigning and delegating tasks as a data steward

Data stewards can assign new tasks to themselves or other data stewards part of the same campaign. They can assign tasks to other data stewards only if tasks are currently unassigned or assigned to themselves.

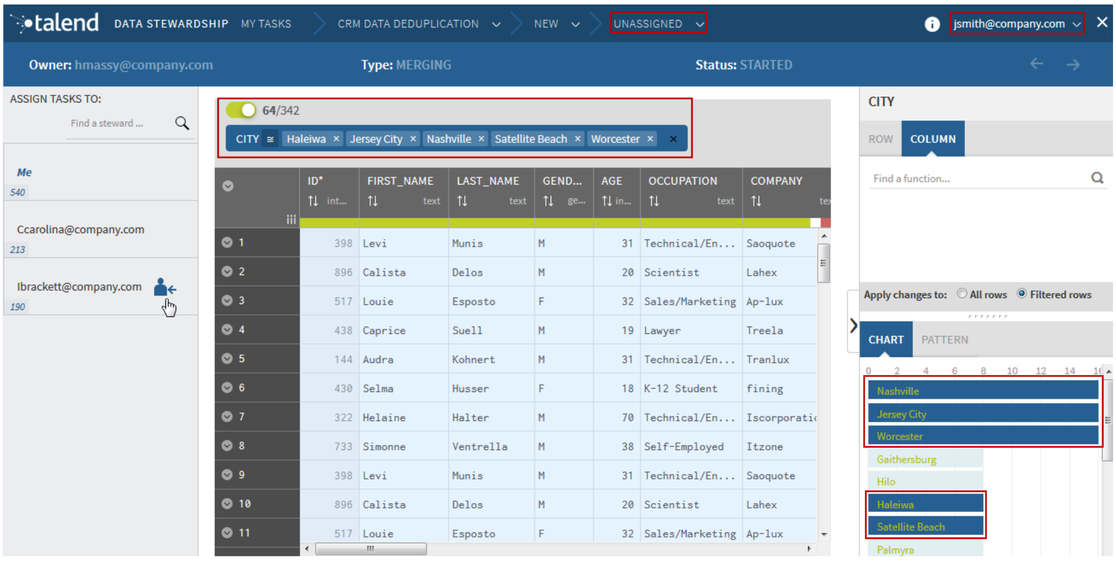
Data stewards can assign or delegate tasks to other data stewards to meet certain needs, for example to improve the distribution of workload across business users or because of expertise on a specific task.

- A campaign owner has loaded the tasks in the campaign using a Talend Job.
- A campaign owner has created the campaign and granted you access to it.

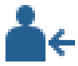
- You have accessed Talend Data Stewardship as a data steward.

Assigning tasks

1. In the **TASKS** page, point to the campaign which holds the tasks you want to assign and click the icon that displays to the right of the campaign name.
   A list of all the unassigned tasks in the campaign is opened.



   You can click **Unassigned** on the menu bar and select **ASSIGNED TO ME** to open the list of the task already assigned to you in the campaign.
2. Use the data quality bar on top of the columns or the charts displayed in the right-hand panel to filter data and list only the tasks you want to assign, or select the tasks manually.

   For further information about filtering tasks with the quality bar or the charts, see the online publication on Talend Help Center (https://help.talend.com).
3. In the left-hand panel, either:
   - Point to the data steward to whom to assign the tasks and click  next to his/her name, or
   - Use the **Assign Tasks To** field to search for the data steward to whom you want to assign the tasks and click .

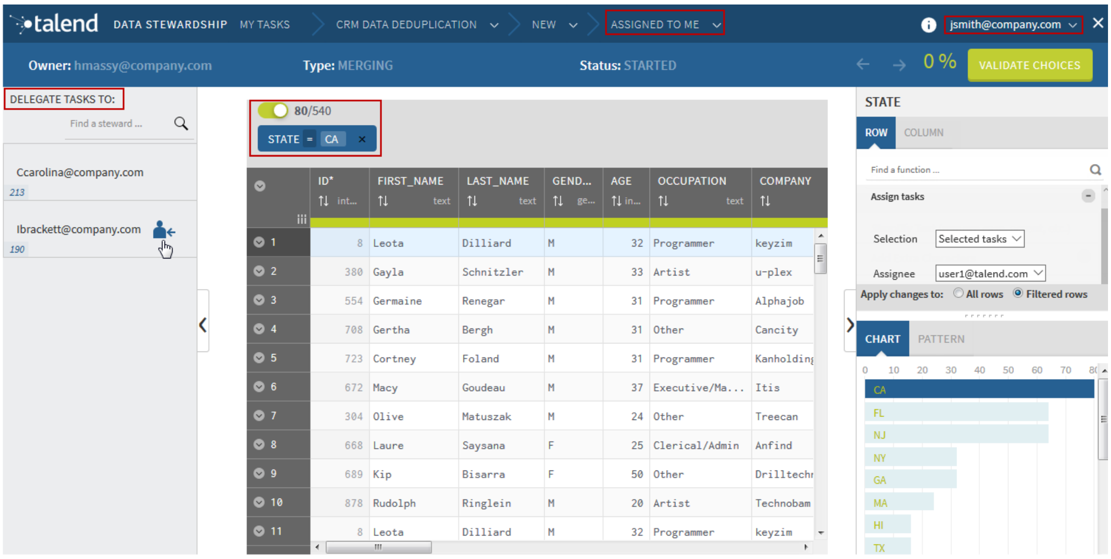
The selected tasks are moved to the dashboard of the data steward. The number of tasks which show under the participant name in the left panel is the number of tasks assigned to him/her within the campaign.

Delegating tasks

The reasons to delegate tasks include the need of certain expertise to handle the tasks, a risk of failing to do the tasks by the appointed time or fast approaching vacations.

Tasks have been assigned to the data stewards.

1. In the **TASKS** page, click the campaign which holds the tasks you want to delegate.



2. Use the chart in the right-hand panel to filter and select a group of the tasks assigned to you and in the left-hand panel, and do one of the following:
   - In the panel to the right, click **ROW** > **Assign tasks** and use the two lists to decide what tasks to delegate to what data steward.
   - Point to the data steward to whom to delegate the tasks and click next to his/her name.
   - Use the **Delegate Tasks To** field to search for the data steward to whom you want to assign the tasks and click next to his/her name.

Handling grouping tasks to decide on relationship among pairs of records

Grouping tasks consist in deciding on a relationship between several records in a group. Once you validate your choice, you transition the task to the second state defined in the workflow.

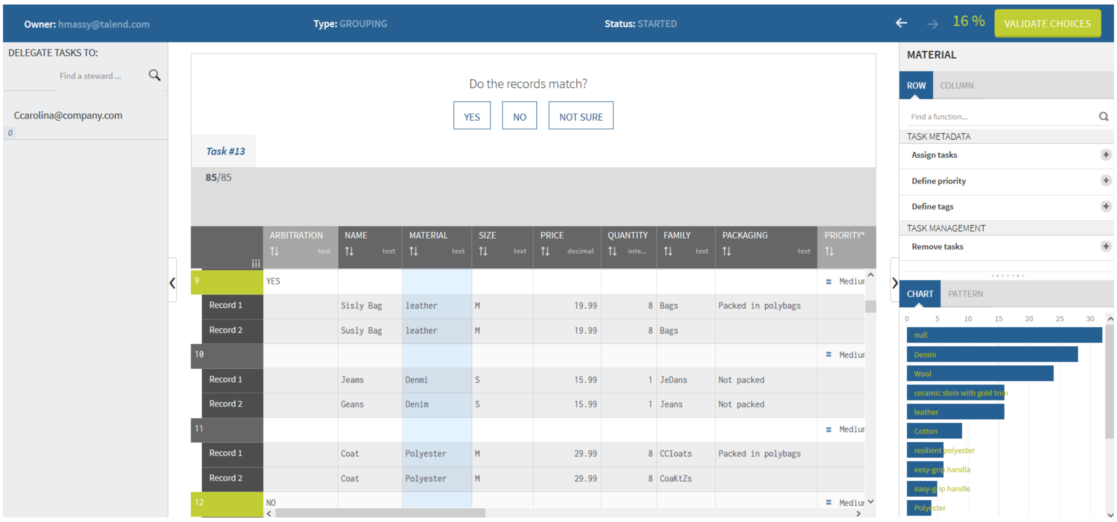
In this example, you need to answer a question to confirm if suspect pairs are duplicates. Once you label the records and validate your choice, a Talend Job retrieves the data from the campaign and uses it in the context of matching data on Spark.

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

  For further information, see Adding a Grouping campaign to identify duplicate pairs on page 40.
- A campaign owner has assigned you tasks in the campaign. Otherwise, you can assign the tasks to yourself.

  For further information, see the online publication about assigning tasks on Talend Help Center (https://help.talend.com).

1. In the **TASKS** page, click the campaign name, `Product deduplication` in this example, to open a list of the tasks assigned to you.



2. Select one task or use the Ctrl/SHIFT key to select multiple tasks and click **Yes**, **NO** or **NOT SURE** to confirm the relationship between data pairs.
   Tasks are tagged with green to show that a decision has been taken on them and your choice is listed in the **ARBITRATION** column.
3. Click **VALIDATE CHOICES** in the top-right corner of the page to validate the choices you have done on the tasks.

Choices are set, data records are resolved, validated and moved from your list.

What to do next:

Use a Talend Job with the **tMatchModel** component which analyzes the data labeled in the `Product deduplication` campaign and uses an algorithm to generate a matching model.

For further information, see the online publication about generating a matching model on Talend Help Center (https://help.talend.com).

Using Integrated Matching with Talend Data Stewardship

Using Integrated Matching with Talend Data Stewardship

An MDM user can set up a project in Talend Studio to create master records from data using integrated matching. Integrated matching groups together similar records and creates a "master" record which is a consolidated version of all records in the group. Records which are not merged automatically are sent to Talend Data Stewardship so that data stewards can intervene on them and resolve the tasks.

The tasks that are listed in Talend Data Stewardship are the result of validating data records in the Staging Area through Talend MDM Web UI. They are listed in a **Merging** campaign created automatically.

When using integrated matching:

- Every time an MDM user runs a validation on a record in the Staging Area and if the record matches an existing group:
  - The record is added to the group in the Staging Area,
  - The task is reopened in Talend Data Stewardship, if it is already closed, and the new duplicate record is added to the task.

- Every time an MDM user updates records that were already integrated as a source or a master record in a task in Talend Data Stewardship, updates are propagated automatically.
- Every time an MDM user redeploys a new data model on MDM, updates are propagated to the data model and campaign in Talend Data Stewardship.

For further information about the integration processes done in Talend MDM, see the Theory into practice: Integrated matching in Talend MDM article published on Talend Help Center (https://help.talend.com).

Defining roles in a campaign created by Integrated Matching

Upon the deployment of the MDM data model in Talend MDM as part of the integrated matching processes, a **Merging** campaign and a data model are automatically created in Talend Data Stewardship. The data model is attached to the campaign to map the MDM data types to the data types in Talend Data Stewardship, and the campaign holds the duplicate records which are not merged automatically.

The names of the campaign and data model are suffixed with `-tmdm`.

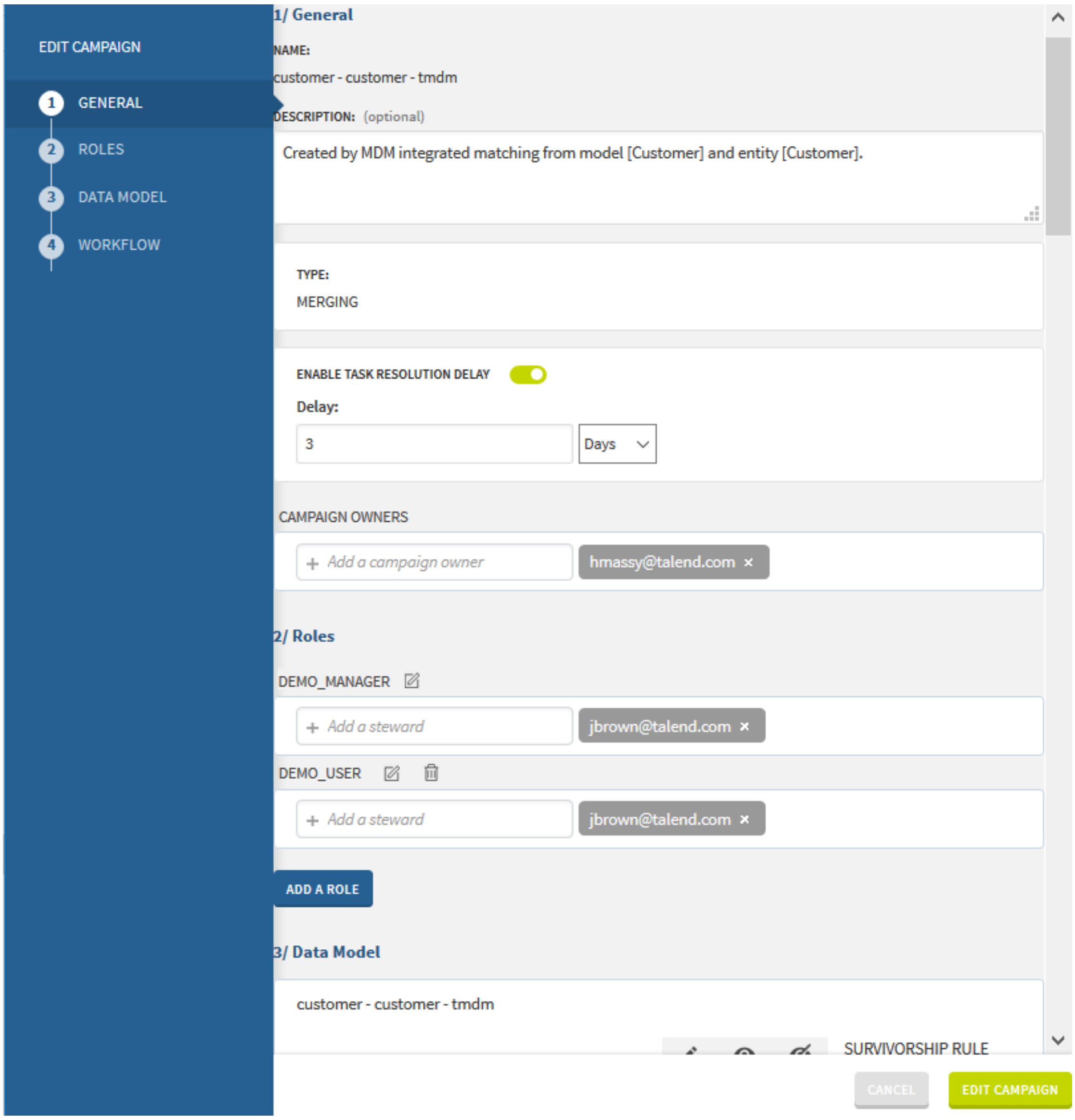
A campaign owner needs now to define what data stewards are part of this **Merging** campaign.

For further information about the integration processes done in Talend MDM to generate **Merging** campaigns, see the Theory into practice: Integrated matching in Talend MDM article published on Talend Help Center (https://help.talend.com).

- 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.

1. Log in to Talend Data Stewardship as a campaign owner which is referenced in the MDM configuration file.

   For further information about configuring the integration with Talend MDM, see the Theory into practice: Integrated matching in Talend MDM article published on Talend Help Center (https://help.talend.com).
2. In the **CAMPAIGNS** page, click the search icon on the top-right corner and enter `tmdm` to filter the list and show only the **Merging** campaigns created by integrated matching.
3. Point to the name of the campaign where you want to add data stewards and click the gear icon which displays to the right to open the campaign metadata.
   The name, type and data model attached to the campaign are read-only.
4. Click **ROLES** in the left-hand panel.
   The roles should match those referenced by the MDM data model. Talend Data Stewardship automatically retrieves associated read/write permissions on each attribute of the data model.



5. 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.
6. Click **EDIT CAMPAIGN** to save your changes.
7. In the home page, click **CAMPAIGNS** and click the name of the campaign to open it and assign the task to the data stewards you have just defined.
   For further information, see Assigning and reassigning tasks as a campaign owner on page 11.

Handling merging tasks created by Integrated Matching

Data stewards with access rights to the **Merging** campaign created automatically in Talend Data Stewardship need to access the merging tasks and manually merge duplicate records into one master record. This will send resolved data back to the Staging Area in Talend MDM Web UI.

In a **Merging** campaign, data stewards can only modify values in the master fields, values in the source fields can not be modified.

Merging data values and validating your modifications transition the task to the Resolved state. A task cannot be validated or even marked as ready as long as it contains at least one invalid value.

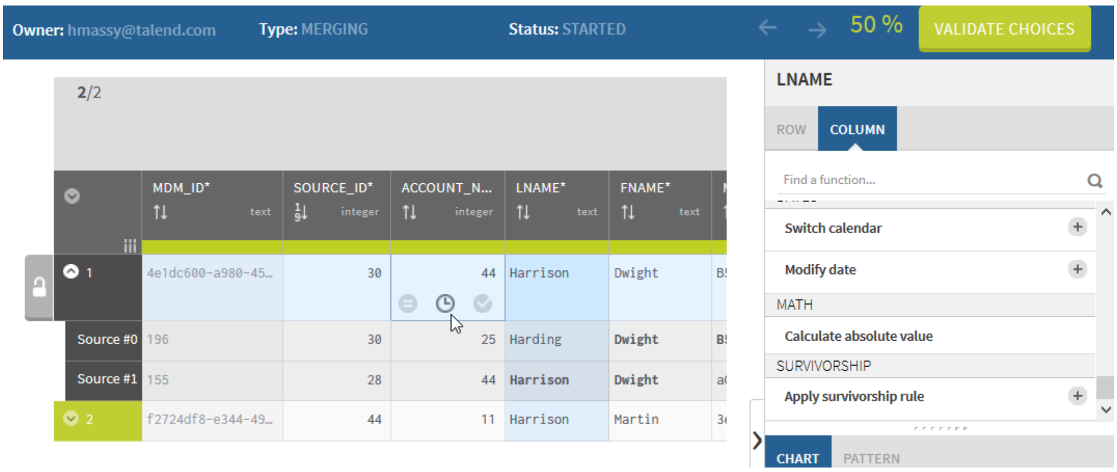
In this example, duplicate records come from Talend MDM as a result of integrated matching processes used to validate customer records against a match rule. Talend Data Stewardship determines initially which attributes of the records to use to create the master record according to the survivorship rules defined in the data model created in Talend MDM and deployed on the server with the **Merging** campaign. However, you may need to manually modify survivorship rules per record attribute or enter completely new values to reach the most accurate and reliable master records.

- An MDM administrator with the campaign owner role has validated the customer records in Talend MDM Web UI to deploy duplicates which need human intervention to Talend Data Stewardship.

  For further information about the integration processes done in Talend MDM to generate **Merging** campaigns, see the Theory into practice: Integrated matching in Talend MDM article published on Talend Help Center (https://help.talend.com).
- A campaign owner has granted you access to the **Merging** campaign.
- A campaign owner has assigned you tasks in the campaign. Otherwise, you can assign the tasks to yourself.

  For further information, see the online publication about assigning tasks on Talend Help Center (https://help.talend.com).
- You have accessed Talend Data Stewardship as a data steward.

1. In the **TASKS** page, click the search icon on the top-right corner and enter `tmdm` to filter and list only the **Merging** campaigns created by integrated matching for which you have access rights.
2. Click the campaign name to open the list of the tasks assigned to you.



3. Use the quality bar on top of each of the columns to filter the data on which you want to work in the **CHARTS** or **PATTERN** views in the right-hand panel.
4. Click the down arrow on the top-left corner to expand all tasks in the list, or click the down arrow of a specific task to expand it.
5. 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 coming from different sources.

    Icons are grayed out when a rule is not applicable on the selected attribute.
  - Set a survivorship rule manually for one attribute of multiple records.
    1. Click a column heading and in the right-hand panel browse to the **Survivorship** section.
    2. Click the button and from the **Survivorship rule** list, select the rule you want to apply to all values in the selected column.
    3. Click **Submit** to select the most common name values and add them to the master records of 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.

6. Repeat the above step to merge records and create master records for all the tasks assigned to you. If a given column has some values which need to be fixed, you can bulk transform them by using the functions listed in the right panel.

   For further information about the functions used to transform data in columns, see the online publication on Talend Help Center (https://help.talend.com)..
7. Click the icon next to the data record you modified to mark the task as ready to be validated. When the lock icon has a red background color, you must first correct the invalid value in the task before being able to mark it as ready to be validated.

   The record is marked with green background and the lock icon is automatically moved to the next record. 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.
8. Click **VALIDATE CHOICES** in the top-right corner of the page to validate the modifications you have done on the records.
   Master records are created and the records which are validated are moved from the list and marked as resolved.

Approved tasks are transitioned to the Resolved state in the workflow and marked as resolved.

Rejected tasks are transitioned back to the initial step in the workflow and marked as new.

A data steward in Talend MDM Web UI needs to run the Staging validation again to take into account the change in the record status.