talend

Connecting the Data-Driven Enterprise

Talend Data Stewardship User Guide

6.4.1

Contents

Copyright

Adapted for 6.4.1. Supersedes previous releases.

Publication date: June 29th, 2017

What is Talend Data Stewardship?

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

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

Licenses and users of Talend Data Stewardship are managed in Talend Administration Center by administrators. Administrators can set or replace a license and also create users and assign them predefined roles.

Campaigns and tasks

The main units of work in Talend Data Stewardship are campaigns and tasks.

What are campaigns?

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

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

Several campaign types are supported including **Arbitration**, **Resolution**, **Grouping** and **Merging**. A campaign contains tasks of homogeneous type. The task type defines what the data steward has to do to resolve tasks.

A campaign can be owned by one or several campaign owners.

What are tasks?

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

When tasks are created, they belong to a campaign and they are initially unresolved and unassigned unless the campaign owner decides to assign them to a specific data steward through the parameters used in the **tDataStewardshipTaskOutput** component.

Tasks have a lifecycle where they pass through different states according to the workflow defined in the campaign. When the current assignee performs some modification on a task, the task is transitioned to the next step in the workflow till it reaches the resolved state. The task type defines what the data steward has to do to resolve it.

Talend Data Stewardship supports several task types including:

- Arbitration tasks: allow authorized data stewards to assign them to a label chosen among a list of arbitration choices defined by the campaign owner. The outcome of an arbitration task is the arbitration choice made by data stewards.
- Resolution tasks: allow authorized data stewards to fix one or several fields of the data record. The outcome of a resolution task is the curated record fixed by data stewards.
- Merging tasks: allow authorized data stewards to merge several potential duplicate source records into one single record (golden record). The outcome of a merging task is the golden record produced by data stewards.

  Source records can come from the same source (database deduplication) or different sources (databases reconciliation).
- Grouping tasks: allow authorized data stewards to validate a relationship between pairs or groups of records. The outcome of a grouping task is the list of records associated to each other.

Built-in roles and workflows

Talend Data Stewardship comes with predefined roles and workflows.

Roles are used to manage user authorization on tasks and functionalities and they are assigned in Talend Administration Center.

Workflows define the steps necessary to resolve the tasks in a campaign. A workflow is composed of states, linked to each other by transitions.

Roles:

- Campaign owners: can be technical or business users:
  - Technical users are in charge of populating campaigns created on the Talend Data Stewardship server with actual data by using Talend Jobs and output governed data to target systems.
  - Business users are in charge of creating campaigns and data models on the Talend Data Stewardship server. They organize, supervise and monitor the tasks they assign to data stewards. Campaign owners can participate as data stewards.
- Data stewards: Business users in charge of resolving data stewardship tasks. they can:
  - classify data by assigning a label chosen among a predefined list of arbitration choices.
  - curate one or several fields of the task record by fixing invalid values or by providing missing values for example.
  - merge several potential duplicate records into one single record.

  They do not need to have particular technical skills.

Workflows:

- Two-state workflow: Defines one initial state and one final state where the newly created tasks are picked up by a given steward, handled and then marked as resolved.
- Three-state workflow: Defines one initial state, a validation state and one final state. The newly created tasks are picked up by a given steward, handled and then forwarded to another steward in charge of approving the task resolution before they are marked as resolved.

Role-based access control

The roles within Talend Data Stewardship provide a way of assigning access rights to different users.

Access control is managed at two levels: from Talend Administration Center and from Talend Data Stewardship:

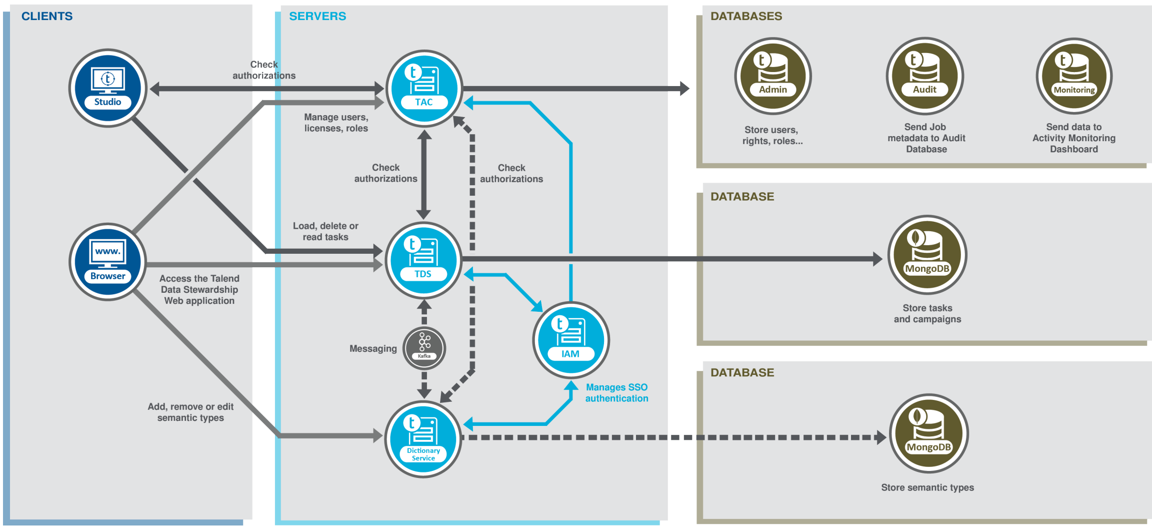
- Roles in Talend Administration Center define if you are a campaign owner and/or a data steward and thus what actions you are entitled to do in Talend Data Stewardship.

  By grouping users into roles in the administration center, administrators can assign and withdraw access rights by assigning or removing memberships to roles.
- Roles within a campaign define which stewards are part of the campaign and which state of the selected workflow they have access to.

It is possible to assign users to multiple roles.

Talend Data Stewardship architecture

This architecture diagram identifies the functional blocks of Talend Data Stewardship, and the interactions among them.



Several functional blocks are defined:

- The **Clients** block, with a Web browser and a Talend Studio.

  From the Web browser, you access the Talend Data Stewardship Web application. This is where campaign owners and data stewards manage campaigns and tasks.

  From Talend Studio, campaign owners can use Jobs to load, retrieve or delete tasks from the campaigns created in the Web application. **tDataStewarshipTaskInput**, **tDataStewardshipTaskOutput**, and **tDataStewardshipTaskDelete** are the three dedicated component to work with Talend Data Stewardship.

  From the command line or a user interface, you can optionally access the Talend Dictionary Service server and add, remove or edit the semantic types used on data in the web application. For further information, see Enriching the semantic types libraries through the command line interface on page 37 and Enriching the semantic types libraries through the user interface on page 23.
- The **Servers** block includes the Talend Data Stewardship application server, connected to Talend Administration Center and optionally to the Talend Dictionary Service server.

  The Talend Identity and Access Management Service is used to enable Single Sign-On.

  Talend Administration Center allows administrators to manage licenses, users and roles. Assigning one or more of the predefined roles to users grants them specific rights to what can they access or perform in Talend Data Stewardship. For further information, see Talend Administration Center User Guide.

Talend Dictionary Service allows campaign owners to manage the semantic categories used on data in the web application. It communicates with the Kafka server.

The Kafka server is used to exchange messages on every single event done in the web application such as modifying a data model, using a data model in a campaign, auditing actions performed on tasks and managing semantic categories using Talend Dictionary Service.

- The **Databases** block contains the databases used with Talend Administration Center and a MongoDB database.

  The Administration database is used to manage user accounts and rights, the Audit database is used to evaluate different aspects of the Jobs implemented in the Studio and the Monitoring database is used to monitor the execution of technical processes and service calls.

  One MongoDB is used to store campaigns and tasks and another MongoDB is used to store semantic types. Nothing is saved directly on your computer.

  However, you can share one MongoDB between Talend Data Stewardship and Talend Dictionary Service. In such a case you need to have two databases on the MongoDB instance.

Administrating Talend Data Stewardship

Creating Data Stewardship users

Talend Administration Center allows you to define Data Stewardship users and assign them predefined roles. This makes the user list accessible from Talend Data Stewardship where campaign owners can assign these users to specific campaigns and tasks.

You can either associate users with only Talend Data Stewardship, or you can associate them with hybrid projects where they can work on campaigns and tasks in Talend Data Stewardship and also work on other project types.

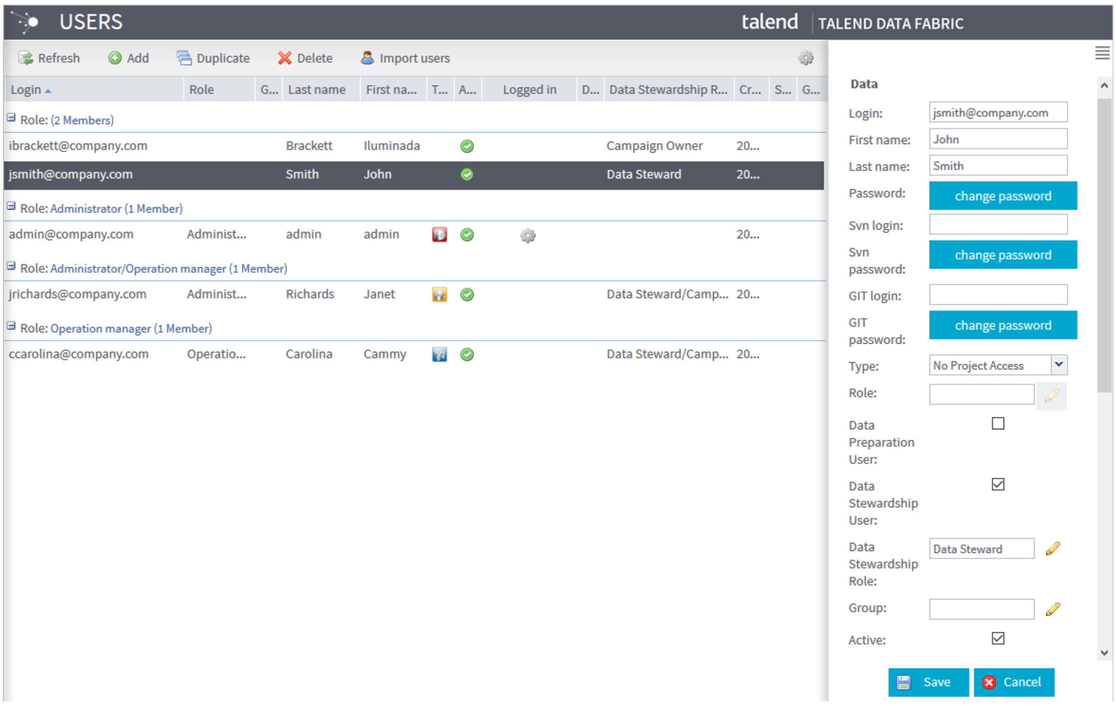
Adding users with Data Stewardship type

This type of user can access Talend Data Stewardship, but he/she does not have access to other projects.

- You have a license to a Talend subscription product.

1. Log in to Talend Administration Center.
2. On the **Users** page, click **Add** to create a new user account.

   The **Data** panel opens to the right, where you need to fill the user's information.



3. Enter the user's name, login (email address) and password for this account.
4. Select the **Data Stewardship User** check box to set this account as a Data Stewardship account.

   If you want to enable the same user to access Talend Data Preparation, select the check box of **Data Preparation User**.

   For further information, see the online publication about creating Talend Data Preparation users on Talend Help Center (https://help.talend.com).
5. From the **Type** list, set the Data Stewardship user type to **No Project access** as this user is not linked to any projects and will only work in Talend Data Stewardship.
6. Click next to the **Data Stewardship Role** to open a dialog box where you can select the check box(es) of the Data Stewardship role(s) you want to assign to the selected user.

   For further information on Data Stewardship predefined roles, see Built-in roles and workflows on page 5.
7. Click **Save** to validate the creation of the new user.

The Data Stewardship user is created and added to the list of users in the **Users** page.

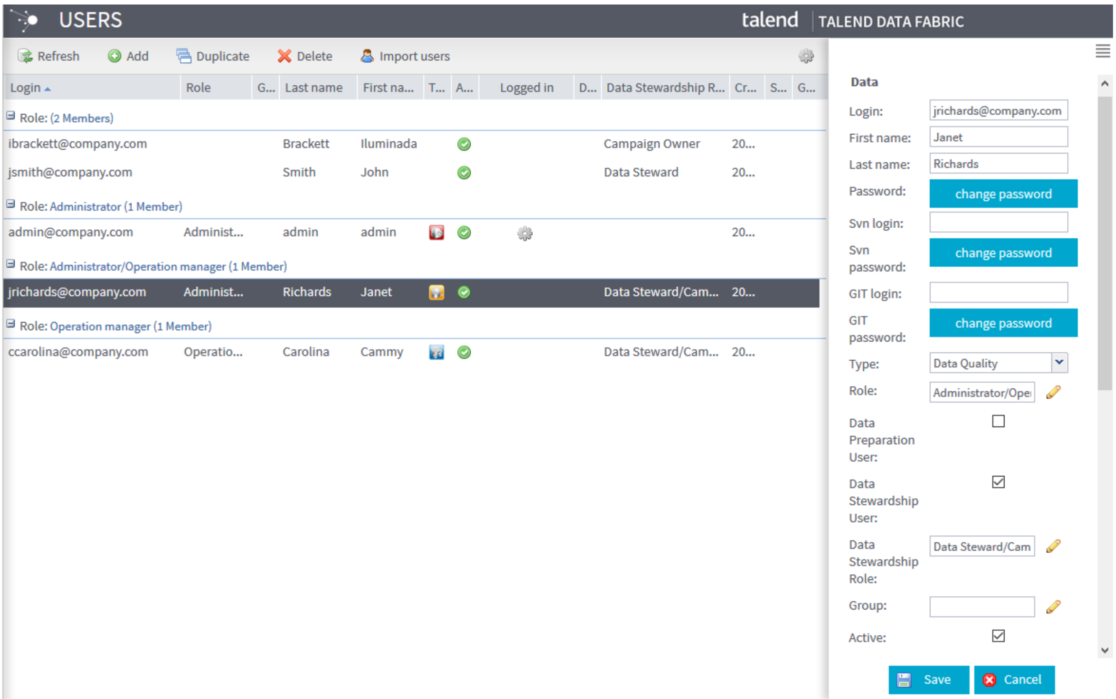
Adding hybrid Data Stewardship users

This type of user can work in Talend Data Stewardship and also on the project type selected from the **Type** list. For example, a hybrid Data Stewardship/Data Quality user can be assigned to a **Data Quality** project type.

- You have a license to a Talend subscription product.

1. Log in to Talend Administration Center.
2. On the **Users** page, click **Add** to create a new user account.

   The **Data** panel opens to the right, where you need to fill the user's information.



3. Enter the user's name, login (email address) and password for this account.
4. From the **Type** list, set the type of the project the Data Stewardship user will be working on.

   The type of accessible projects depends on the license you have, namely **Data Integration/ESB**, **Data Quality** or **Master Data Management**.

   If you enabled the **Role Mapping** option in the **SSO** node of the **Configuration** page, these fields might be automatically filled. For more information about enabling SSO, see the Talend Administration Center User Guide.
5. Click  next to the **Role** field to open a dialog box where you can select from the list the check box(es) of the administration role(s) you want to assign to the selected user.
6. Select the **Data Stewardship User** check box to set this account as a Data Stewardship account.

   If you want to enable the same user to access Talend Data Preparation, select the check box of **Data Preparation User**.

   For further information, see the online publication about creating Talend Data Preparation users on Talend Help Center (https://help.talend.com).
7. Click  next to the **Data Stewardship Role** to open a dialog box where you can select the check box(es) of the Data Stewardship role(s) you want to assign to the selected user.

   For further information on Data Stewardship predefined roles, see Built-in roles and workflows on page 5.
8. Click **Save** to create the user and list it in the **Users** page.

   To create users via the MetaServlet application, use the `createUser` command and define the `tds` and `tdsRoles` arguments. For more information about parameters used in metaServlet, see the Talend Administration Center User Guide.

   For more information about how to add a user with LDAP, see the Talend Administration Center User Guide.

Talend Data Stewardship start and stop sequences

In order to start and stop Talend Data Stewardship and its dependencies, you need to follow a certain sequence.

The availability of Talend Dictionary Service depends on your subscription level.

To start Talend Data Stewardship and its dependencies, the order is the following:

1. Apache Zookeeper
2. Kafka
3. MongoDB
4. Talend Administration Center
5. Talend Identity and Access Management Service
6. Talend Dictionary Service
7. Talend Data Stewardship

The stop sequence is the reversed version of the start sequence.

Backing up a Talend Data Stewardship instance

Backing up Talend Data Stewardship on regular basis is important to recover from a data loss scenario or any other causes of data corruption or deletion.

If you want to have a copy of a Talend Data Stewardship instance, you need to backup the MongoDB database, the folders containing your data, the configuration files and the logs.

This procedure backup MongoDB with mongodump, but you can use different backup methods. For further information, see MongoDB Backup Methods.

If you have been using Talend Dictionary Service, shut it down before starting the backup to ensure having exact copies of the index files.

1. Start your MongoDB.
2. Stop your Talend Data Stewardship instance.
   You can execute the same backup command while Talend Data Stewardship is running. However, a hot backup may cause poor performance.
3. To back up your MongoDB, open a command prompt window and execute the following command:
   ```
   <mongo_path>/bin/mongodump -h
   <source_mongodb_host>:<source_mongo_port> -u <mongo_tds_username>
   -p <mongo_tds_password> -d <mongo_tds_database> -o
   <backup_location> --oplog
   ```
4. To back up any configuration you might have done, make a copy of the `<path_to_installation_folder>/tds/apache-tomcat/conf` and the `<path_to_installation_folder>/kafka/conf` folders and save them to a secure place.
5. To back up your logs, copy the log folder at `<path_to_installation_folder>/tds/apache-tomcat/logs` and save it to a place of your choice.
6. If you have installed and used Talend Dictionary Service to create custom semantic types or update the existing ones, complete the following steps:
   a) To back up any changes made to the predefined semantic types, copy the data folder at `<path_to_installation_folder>/tds/apache-tomcat/data` and save it to a place of your choice.
   b) Back up the Talend Dictionary Service server. For further information, see Talend Dictionary Service back up and recovery on page 11.

You have now a backup of the Talend Data Stewardship instance.

Restoring a Talend Data Stewardship instance

After backing up Talend Data Stewardship, you can recover the database and the configuration files at any time.

In order to restore your backup files, you need to paste them to the appropriate location in the Talend Data Stewardship installation folder.

If you use Talend Dictionary Service, shut it down before starting the operation to ensure having exact copies of the index files.

1. Start your MongoDB.
2. Stop your Talend Data Stewardship instance.
   You can restore Talend Data Stewardship while it is running. However, a hot restore may cause poor performance.
3. To restore your MongoDB, open a command prompt window and execute the following command:
   ```
   <mongo_path>/bin/mongorestore -h <mongo_host>:<mongo_port>
   -u <mongo_tds_username> -p <mongo_tds_password> -d
   <mongo_tds_database> --oplogReplay <backup_location>
   ```
4. To restore your configuration, replace the content of the `<path_to_installation_folder>/tds/apache-tomcat/conf` and `<path_to_installation_folder>/kafka/conf` folders with your backed up copies.
5. To restore your log file, replace the content of the `<path_to_installation_folder>/tds/apache-tomcat/logs` folder with your backed up copy.
6. If you have installed and used Talend Dictionary Service to create custom semantic types, or update the existing ones, complete the following steps:
   a) To restore your custom semantic types, replace the content of `<path_to_installation_folder>/tds/apache-tomcat/data` with your backed up copy.
   b) Restore the Talend Dictionary Service server. For further information, see Talend Dictionary Service back up and recovery on page 11.

Talend Data Stewardship now is an exact copy of the instance you backed up.

Talend Dictionary Service back up and recovery

To recover from a data loss scenario or any other causes of data corruption or deletion, you are advised to perform back-ups of Talend Dictionary Service on regular basis and store the database and files in a secure place. However, the availability of the Dictionary service depends on the license you have.

Talend Dictionary Service stores all the predefined semantic types used in Talend Data Stewardship. It also stores all new types entered by users and all the modifications done on existing semantic entries.

To back up a Dictionary service instance, you need to back up the MongoDB database and the changes made to the predefined semantic types.

Backing up Talend Dictionary Service

Backing up a Talend Dictionary Service instance includes backing up the MongoDB which stores all the semantic entries.

Before starting the backup procedure, you are advised to copy the dump file at `<path_to_installation_folder>/dq_dict/database/dump` and save it in a place of your choice to keep a copy of the initial data. This dump file has the by-default semantic types.

1. Stop your Talend Dictionary Service instance.

You can do the backup while the instance is running, but you are advised to choose a period of low activity.

2. Start MongoDB.
3. Open a command prompt window and execute the following commands to backup the current Dictionary service including the logs to a *dump* folder:
   On Windows:

   - <path_to_installation_folder>\dq_dict\database
   - semantic-dictionary-export.bat

   On Linux:

   ```
   <path_to_installation_folder>/dq_dict/database/semantic-dictionary-export.sh
   ```
4. To back up the log and configuration files, make a copy of the `<path_to_installation_folder>/dq_dict/apache-tomcat/logs` and the `<path_to_installation_folder>/dq_dict/apache-tomcat/conf` folders and save them to a secure place.

The backup of the current Dictionary service is completed and the file is saved at

`<path_to_installation_folder>/dq_dict/database/dump`.

However, you will also need to back up the changes made to the predefined semantic types using Talend Dictionary Service and stored in the *tds* folder. For further information, see Backing up a Talend Data Stewardship instance on page 10.

Restoring Talend Dictionary Service

Once you back up Talend Dictionary Service, you can recover the data at any time and have an exact copy of the backed up instance.

1. Stop your Talend Dictionary Service instance.
   You can do the backup while the instance is running, but you are advised to choose a period of low activity.
2. Launch MongoDB.
3. Open a command prompt window and execute the below command to delete the database in the current instance of the Dictionary service:

   ```
   <path_to_installation_folder>/mongodb/bin/mongo <dbname>  -u <userName> -p
    <userPassword> --eval "db.dropDatabase()"
   ```

   If you installed Talend Data Stewardship with the Dictionary service using Talend Installer, the command reads as follows:

   ```
   <path_to_installation_folder>/mongodb/bin/mongo dqdict -u dqdict-user -p duser --eval
    "db.dropDatabase()"
   ```
4. Replace the *dump* folder of the current instance stored at `<path_to_installation_folder>/dq_dict/database/` with the *dump* folder you get from the back up procedure.

5. From a command prompt window, execute the command to import the backup directory of the Dictionary service named *dump*. Use .bat or .sh according to your system, for example

   semantic-dictionary-import.bat

6. To restore the logs and your configuration, replace the content of the <path_to_installation_folder>/dq_dict/apache-tomcat/logs and the <path_to_installation_folder>/dq_dict/apache-tomcat/conf folders with your backed up copies.

The Dictionary service instance is recovered.

Republishing semantic types to Talend Data Stewardship and Talend Data Preparation

As an administrator, you can overwrite all local semantic types of Talend Data Stewardship and/or Talend Data Preparation with the last content published on the Talend Dictionary Service server and stored in MongoDB.

Republishing semantic types becomes necessary when Talend Data Stewardship and/or Talend Data Preparation are out of synchronization because a maintenance operation had to be performed on the server, to restore a previous image for instance. In that case, you need to republish the last content from the Talend Dictionary Service server.

- Make sure the service consumers, Talend Data Stewardship and/or Talend Data Preparation, and all their dependencies are up and running.

Open a command prompt window and in the <Dictionary_Service_Path>/dq_dict/command-line folder, execute the following command according to your operating system:

- category_manager.bat -s -login MY_USER_ID -password MY_USER_PASSWORD
- category_manager.sh -s -login MY_USER_ID -password MY_USER_PASSWORD

This command must be executed on the machine where the command line for Dictionary Service is installed.

All semantic types in local registries in both Talend Data Stewardship and Talend Data Preparationare synchronized with what is published in MongoDB.

Adding and managing data models

Adding data models

Campaign owners should create data models to determine the structure of the data to be managed in a campaign. Talend Data Stewardship validates data accordingly when data is loaded in the campaign.

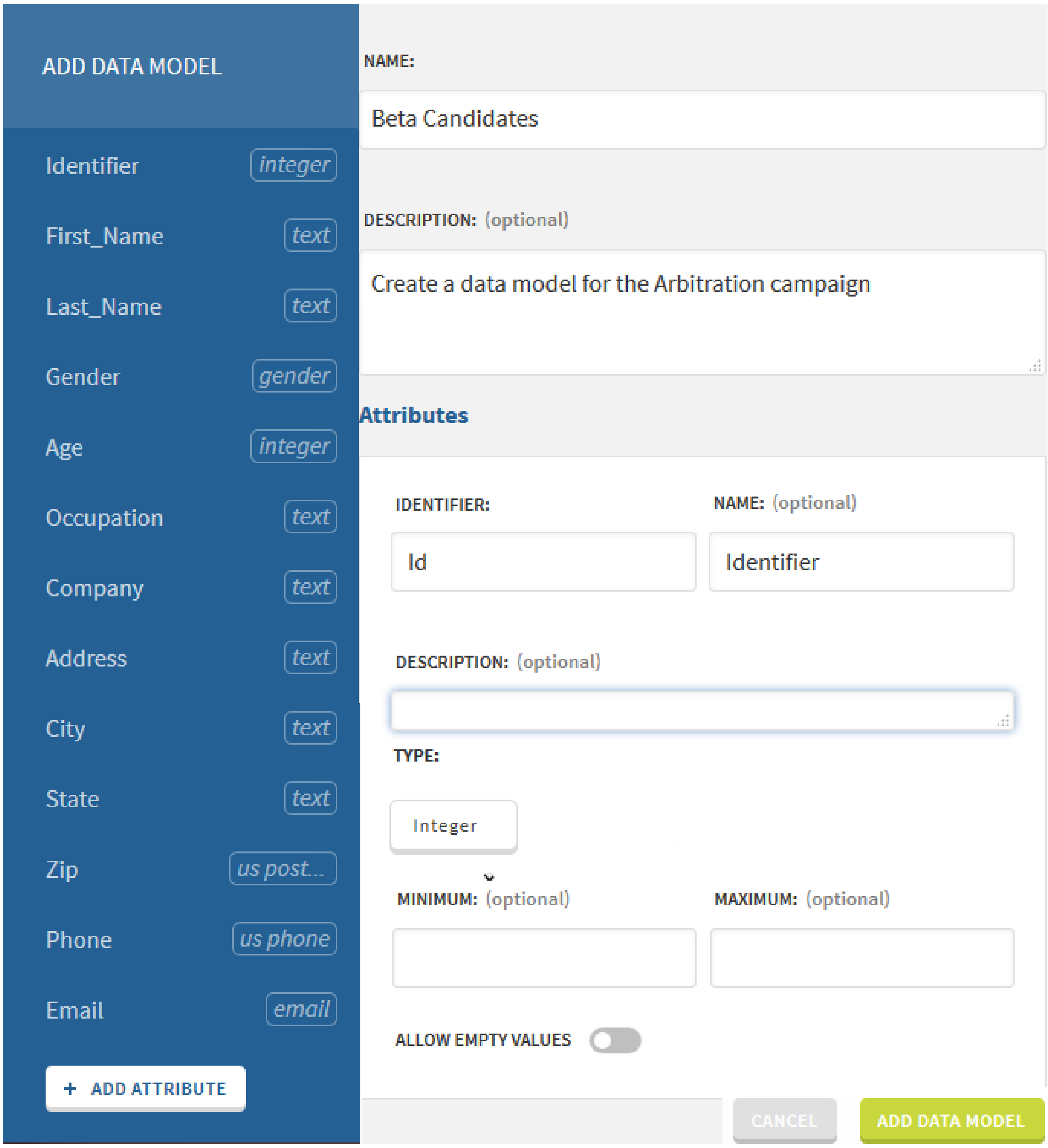
- You have been assigned a campaign owner role in Talend Administration Center.
- You have accessed Talend Data Stewardship as a campaign owner.

Defining a data model for the Arbitration Campaign

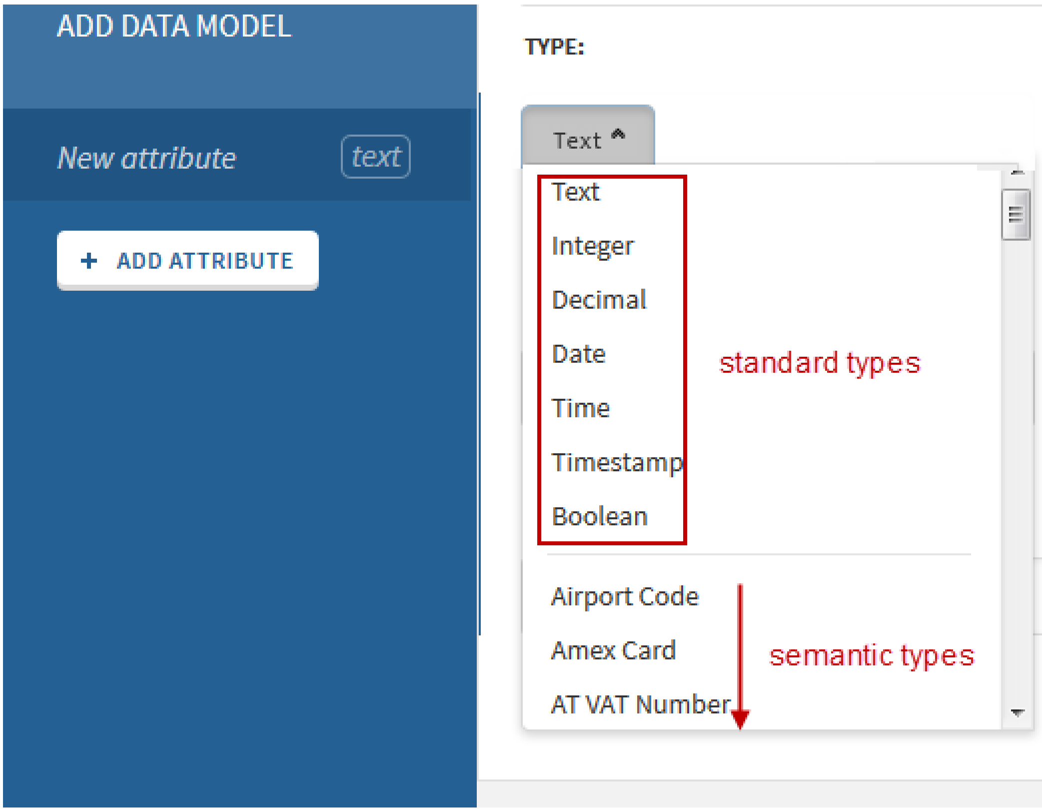
In this example, you create a data model to determine the structure of the data to be managed in the `Beta Candidates` campaign which you create to allow data stewards to select the appropriate candidates of a software beta testers.

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.

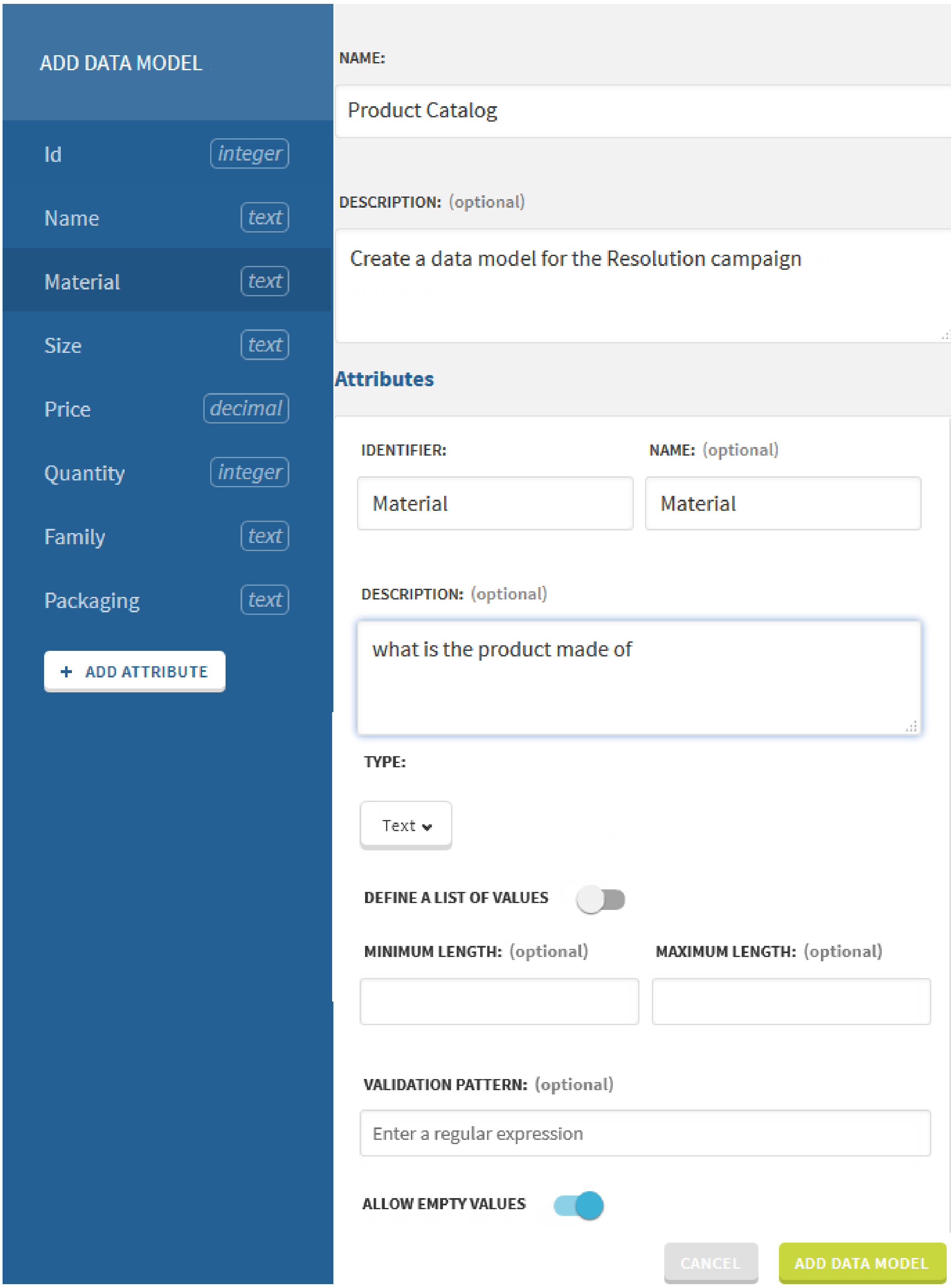
4. If needed, click the switch next to **ALLOW EMPTY VALUES** to disable the load of empty fields to Talend Data Stewardship. This option is enabled by default.
5. Click **ADD ATTRIBUTE** in the left panel and repeat the above steps to create all the columns you need in the data model.
   The columns 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 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.

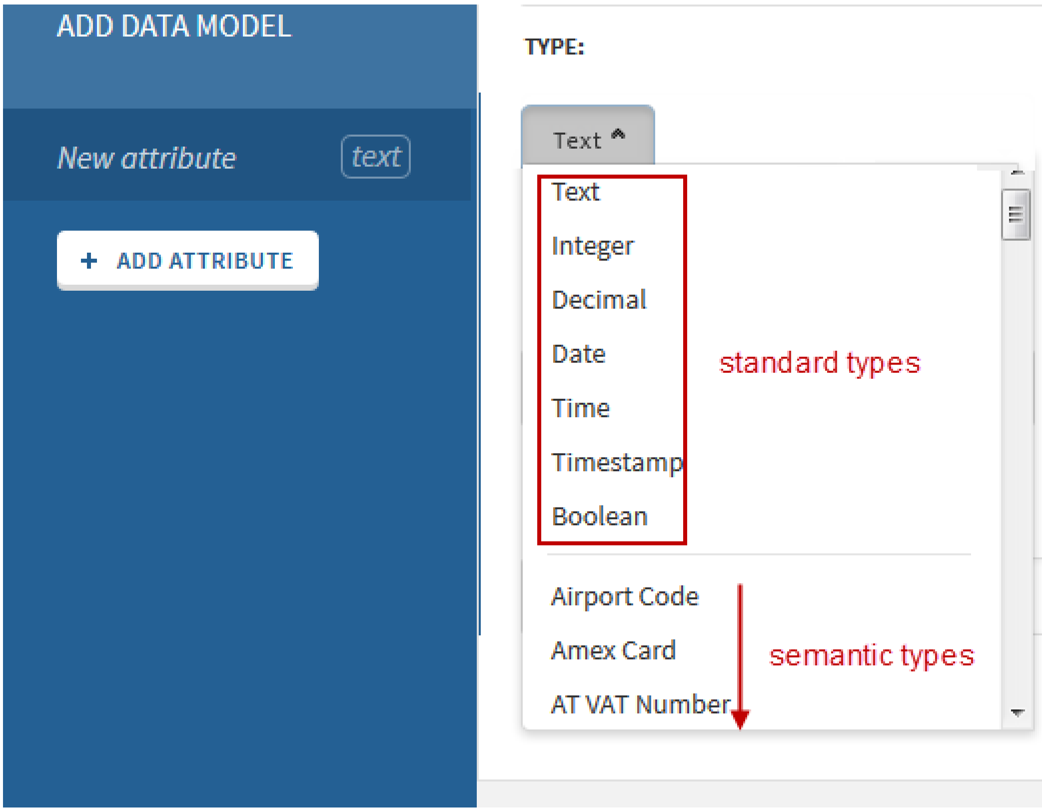


3. In the **Attributes** section, define the columns you want to have in the data model as the following:

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.

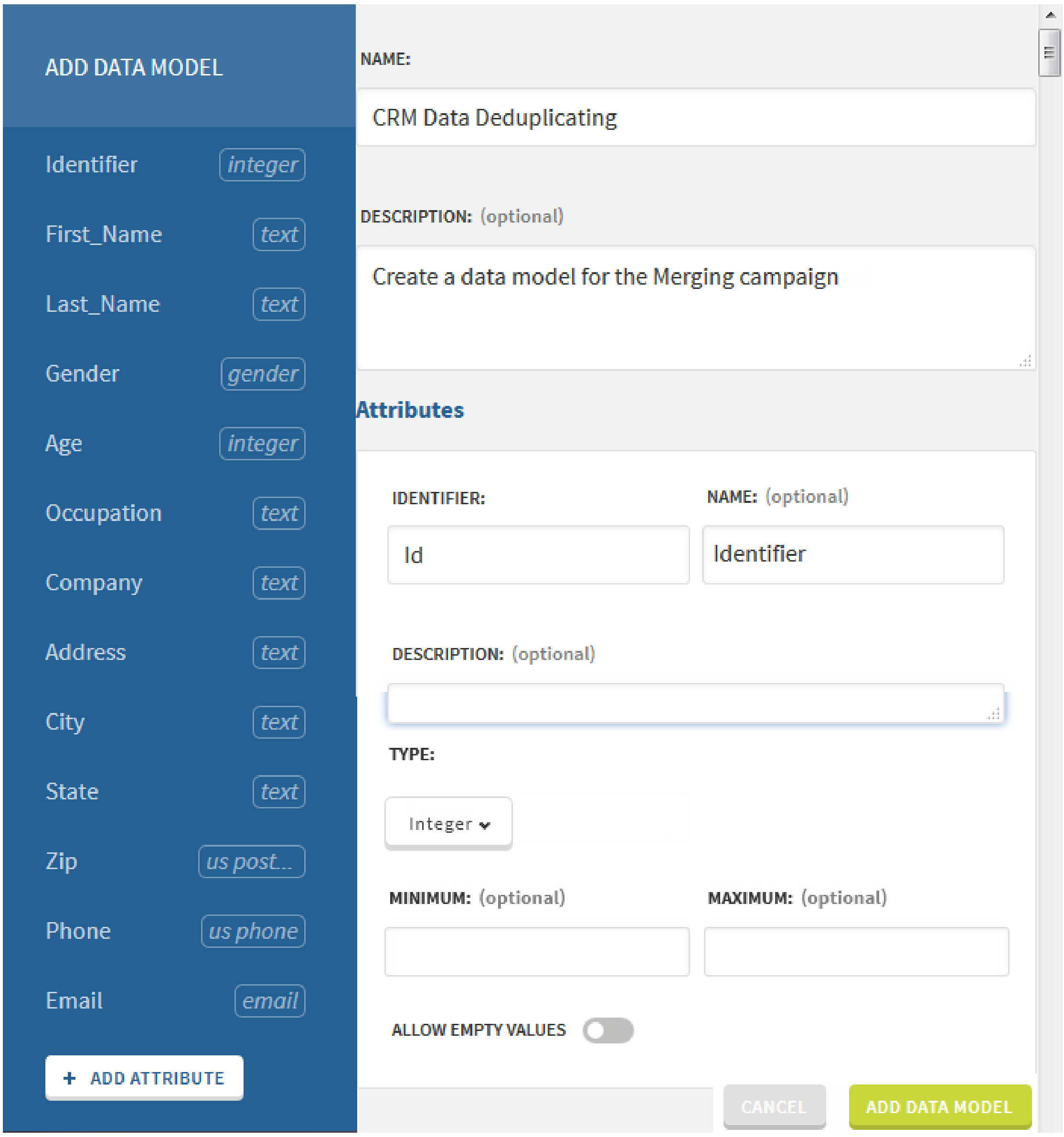
4. If needed, click the switch next to **ALLOW EMPTY VALUES** to disable the load of empty fields to Talend Data Stewardship. This option is enabled by default.
5. Click **ADD ATTRIBUTE** in the left panel and repeat the above steps to create all the columns you need in the data model.
   The columns 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 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.

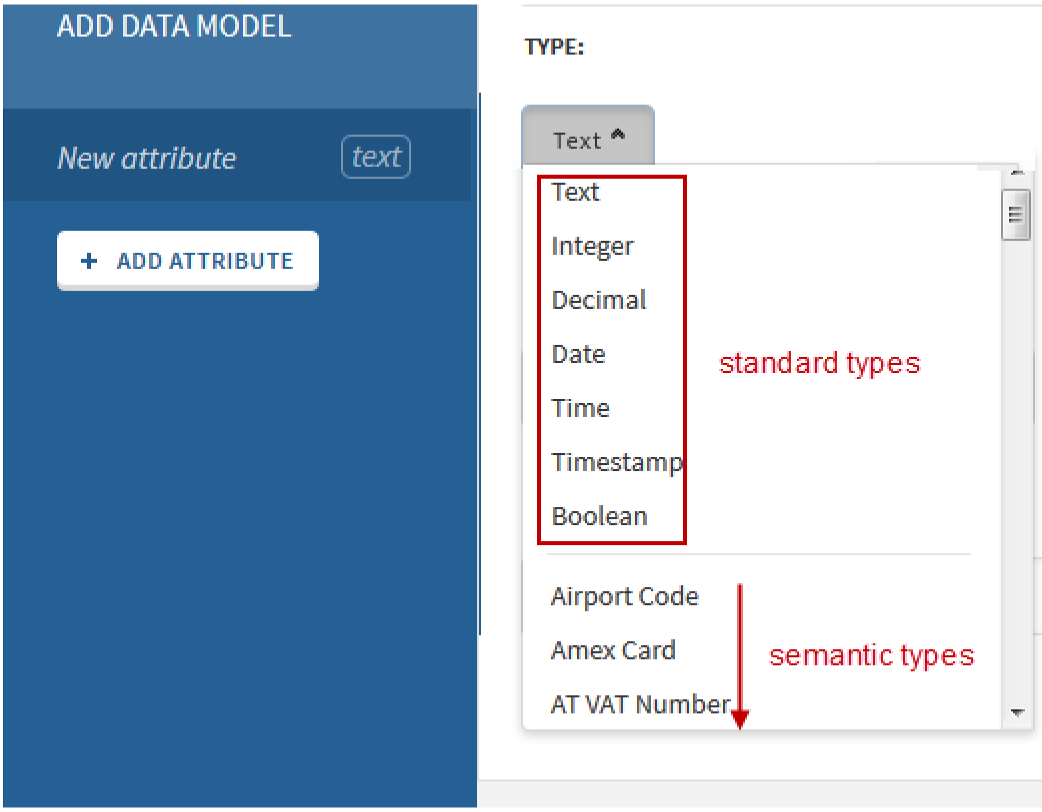


3. In the **Attributes** section, define the columns you want to have in the data model as the following:

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.

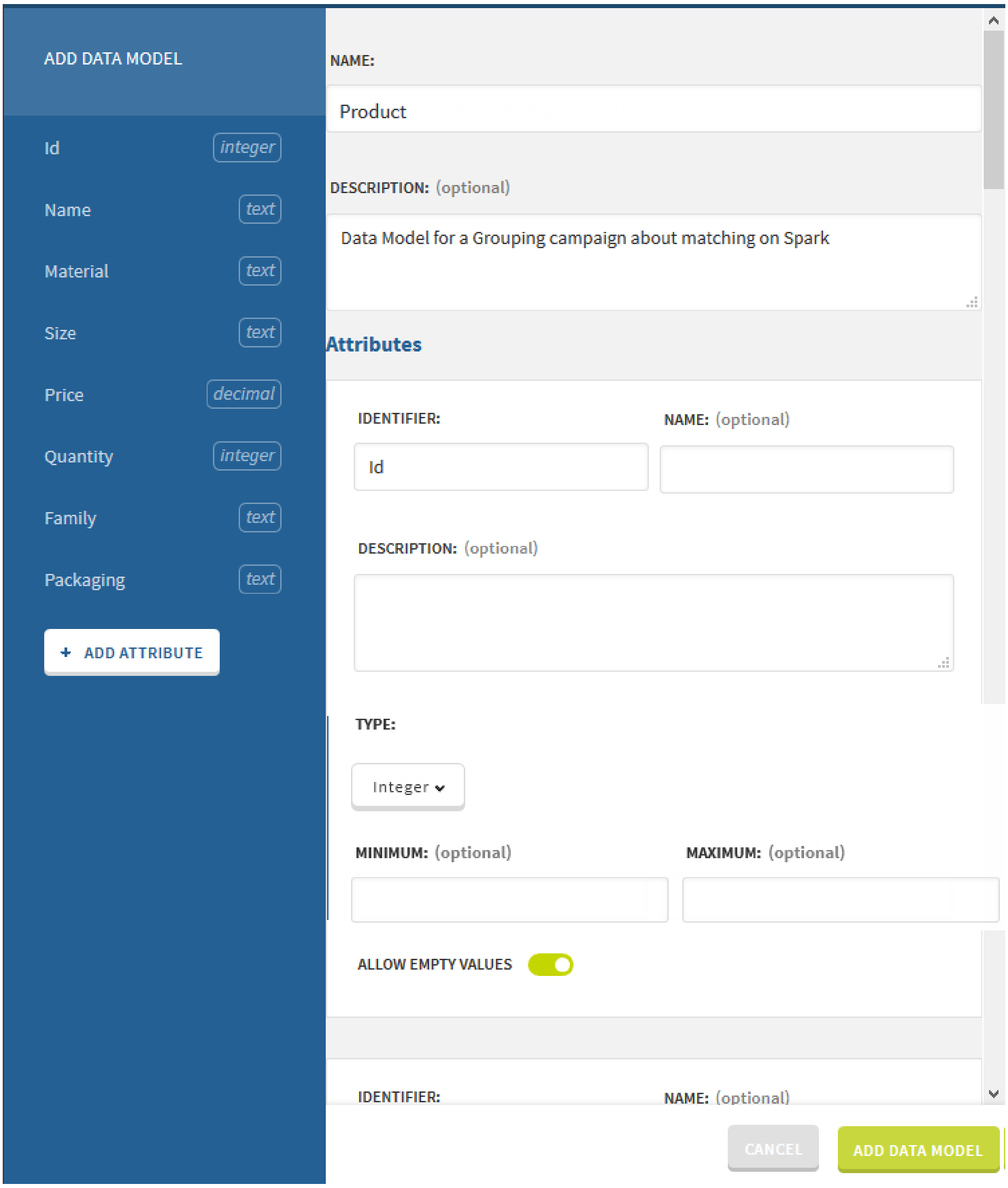
4. If needed, click the switch next to **ALLOW EMPTY VALUES** to disable the load of empty fields to Talend Data Stewardship. This option is enabled by default.
5. Click **ADD ATTRIBUTE** in the left panel and repeat the above steps to create all the columns you need in the data model.
   The columns 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 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.

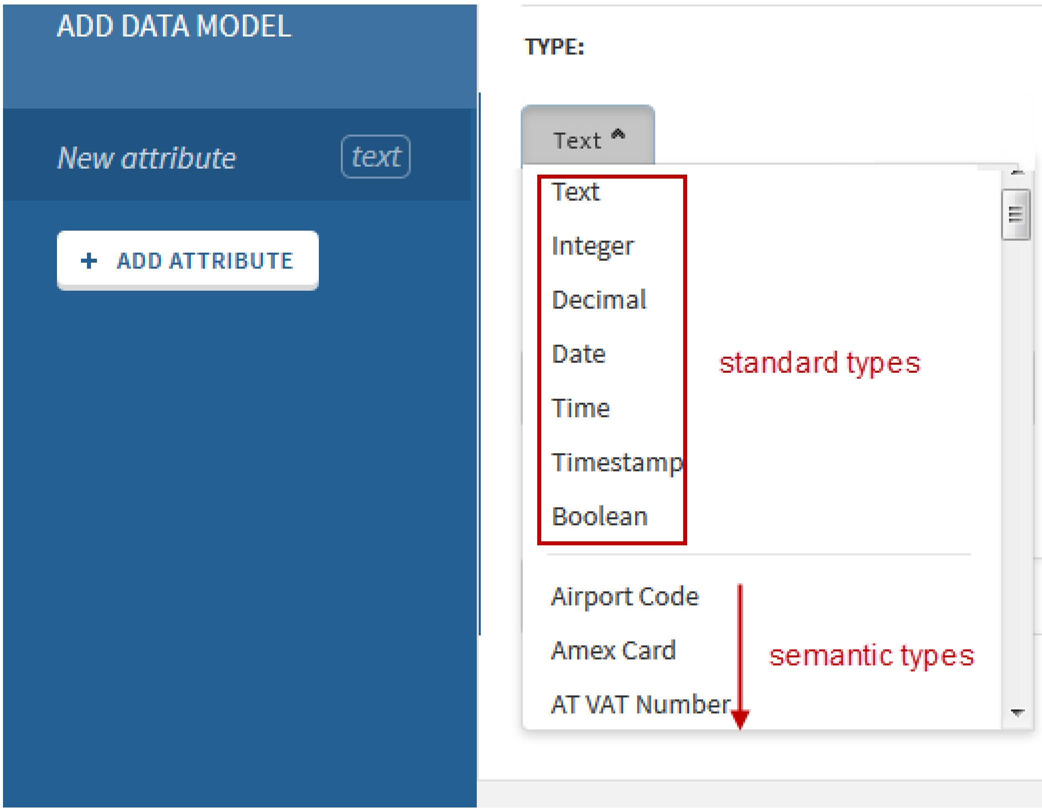


3. In the **Attributes** section, define the columns you want to have in the data model as the following:

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.

Managing data models

Once data models are created and listed in Talend Data Stewardship, campaign owners can perform some basic management procedures on them including editing or deleting data models.

Editing a data model

Campaign owners can update data models whether they are attached to campaigns or not.

When updating a data model used by one or more campaigns, the campaigns are updated as follows:

- Campaigns are temporarily disabled,
- Campaign metadata is updated,
- Tasks are updated, attributes are renamed and the tasks are revalidated,
- Campaigns are enabled.

- The data model has been defined in Talend Data Stewardship.
- You have accessed Talend Data Stewardship as a campaign owner.

1. In the home page, click **DATA MODELS** and click the name of the data model you want to open.
   Campaigns which use the data model are temporarily disabled.
2. Modify the data model metadata or attribute definition, maximum/minimum value or the pattern used on the attribute.
   You can also add new attributes to the data model. However, when you change an attribute identifier, you need to modify accordingly the Talend Job you use to communicate with the campaign and write, read or delete tasks.
3. Click **EDIT DATA MODEL** at the bottom of the page.
   Campaign metadata is updated, attributes are renamed and the tasks listed in the campaign which are not resolved yet are updated and revalidated.

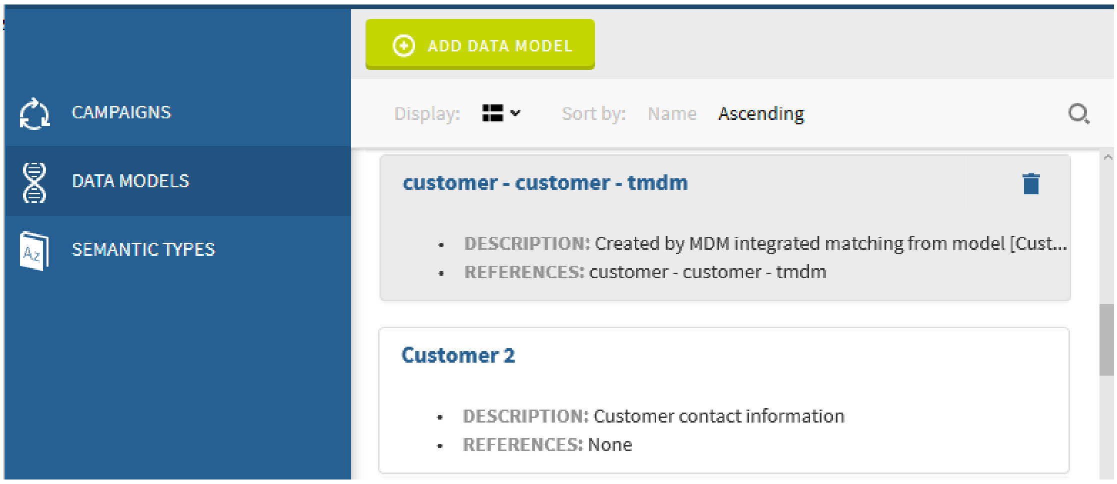
The data model is updated according to the modifications done and the campaigns using the data model are enabled.

Deleting a data model

You can delete any of the data models listed in Talend Data Stewardship for which you have access rights. However, you can not delete a data model which is used in a campaign.

- The data model has been defined in Talend Data Stewardship.
- The data model is not used in any campaign.
- You have accessed Talend Data Stewardship as a campaign owner.

1. In the home page, click **DATA MODELS** to open the list of the data models for which you have access rights.



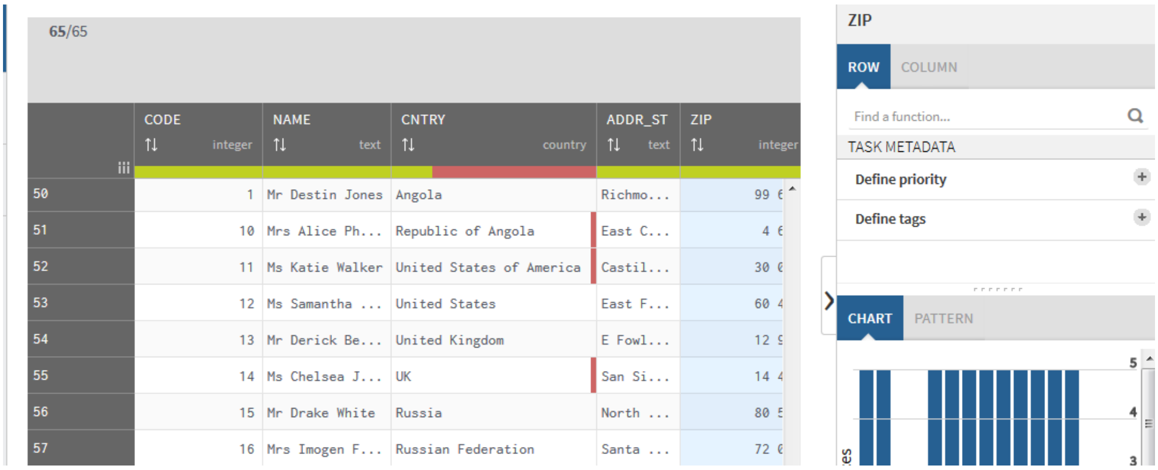
2. Point to the data model you want to delete to display a trash can to the right of the name.
3. Click the trash can icon and confirm the operation when prompted.
4. Click **DELETE**.

The data model is deleted from the list.

Enriching the semantic types libraries through the user interface

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.

When campaign owners define the structure of the data to be managed in a campaign, they can select from a predefined list the semantic type for each attribute. Then when they load data into Talend Data Stewardship, an internal validation of the schema type is performed and data is displayed as valid or invalid accordingly.



For example, the list of entries included by default in Talend Data Stewardship under countries does not include `Republic of Angola`, `United States of America` and `UK`. As a result, such entries are considered invalid country names when loaded to Talend Data Stewardship. For further information about the semantic types present by default, see Predefined standard and semantic types on page 34.

But, you can go further and create your own semantic types, as well as updating or deleting the existing ones, so that your experience with Talend Data Stewardship speaks your business language. You can do all these management options either through a user interface integrated in Talend Data Stewardship or through the command line interface.

When you create semantic types, you can decide either to use them for data validation or to use them for data discovery:

- Data validation matches data against semantic types and marks data as valid or not valid.
- Data discovery explores the semantic categories and query complex semantic relationships in the data you analyze and outputs the matching results to show the most relevant concepts.

Talend Data Stewardship uses the semantic types only for validation as no data discovery is done its side.

For further information about using the command line to manage semantic types, see Enriching the semantic types libraries through the command line interface on page 37.

Talend Dictionary Service stores all the semantic libraries used in various Talend products, including Talend Data Stewardship. All the changes that you make on the Talend Dictionary Service server are instantly available in Talend Data Stewardship. The availability of Talend Dictionary Service depends on the license you have.

On the server, the semantic types are divided into several categories:

- The `Dictionary` type, based on a closed list of values.
- The `Regular expression` type that compares your data against a preselected regular expression.

- The `Compound type` that compares your data against several semantic types referenced in the compound type.

To enable the interaction between Talend Dictionary Service and Talend Data Stewardship, you must fulfill the following prerequisites:

- Talend Dictionary Service is installed and running.
- Talend Administration Center is installed and running.
- Your license in Talend Administration Center includes Data Quality.
- The role assigned to you in Talend Administration Center is either **Designer** or **Operation manager**, and your user type is **Data management** or above.
- The **Data Stewardship User** check box is selected for your user in the administration center with any of the two possible roles set in the **Data Stewardship Role** field.
- In the <install_folder>\tds\apache-tomcat\conf\data-stewardship.properties file, the `dataquality.dictionaryservice.enable` property is set as `true`.

Adding a new regular expression-based semantic type

You can create a semantic type based on a regular expression in Talend Dictionary Service and add it to the list of recognized data types in Talend Data Stewardship.

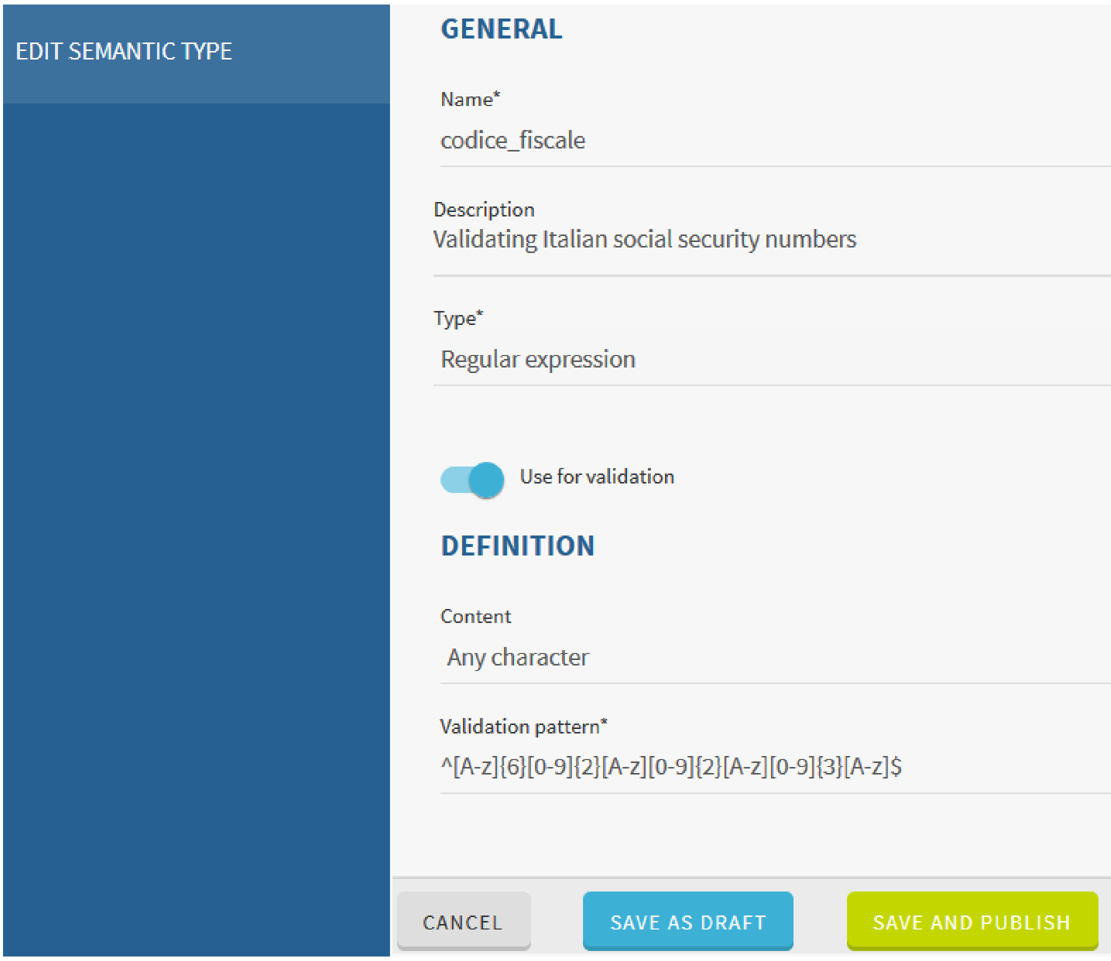
In Talend Dictionary Service, not every type of data can currently be matched with and validated against one of the predefined semantic types. Italian social security numbers, also known as codice fiscale, are currently not recognized for example.

Let's say that you work for an Italian company, only dealing with Italian customers. In this example, you need to intervene and manage some customer data, such as their names, email address, or their social security number.

When defining the data model in Talend Data Stewardship, you will be obliged to set the semantic type for the column containing the social security number to `text` as there is no predefined semantic type for Italian social security number. This is a bit disappointing and you would like to create a more specific category in order to match this type of data: a `codice_fiscale` semantic type in this case.

You will create this new semantic type in Talend Dictionary Service, and it will be automatically available in Talend Data Stewardship so that your data can be matched with and validated against a proper type.

1. In the home page, click **SEMANTIC TYPES** > **ADD SEMANTIC TYPES**.



2. Enter a name and a description for the new semantic type.
3. Select the semantic type from the **Type** list.
4. Leave the switch next to **Use for validation** activated.
   Talend Data Stewardship uses semantic types only for validation.
5. From the **Content** list, select the type of content you want to validate.
   This option helps to optimize performance. Only the data which matches the selected type is validated

Option	Description
Any character	The complete string is validated against the regular expression
Alphabetic	Strings that contain alphabetical characters and no numeric character are validated against the regular expression
Numeric	Strings that contain numeric characters and no alphabetic character are validated against the regular expression

6. Enter the regular expression syntax in the **Validation pattern** field.

   This regular expression is designed to match the Italian codice fiscale, which is an alphanumeric code of 16 characters.
7. Click **SAVE AND PUBLISH** to send the semantic type to the Talend Dictionary Service server and make it available to be used by the system.
   Clicking **SAVE AS DRAFT** stores the new type on the server without propagating it to the system. The new type is not usable unless it is published. For a use case of this option, let's say that you

have new semantic types to deploy as part of a new project. You can prepare the work by creating the semantic types and save them as draft before the go-live of the project, and can deploy the semantic types only the day of go-live.

8. Go back to Talend Data Stewardship and create the data model for the Italian customers data.

   The new semantic category `codice_fiscale` is available now in the list of semantic types and you can set it for the column containing the social security number.

When you load the customer data to Talend Data Stewardship, data is now matched with and validated against the `codice_fiscale` semantic type, that you created in Talend Dictionary Service.

Adding a new dictionary-based semantic type

You can create a semantic type based on a dictionary in Talend Dictionary Service and add it to the list of recognized data types in Talend Data Stewardship. However, duplicate values are not allowed in a dictionary-based semantic type as they are useless and can slow down the process.

In Talend Data Stewardship, not every type of data can currently be matched with one of the predefined semantic types. The counties of United Kingdom for example, are currently not recognized as such.

Let's say that you work for a British company, with customers only residing in the United Kingdom. In this example, you need to intervene and manage some customer data, such as their names, email address, or the county they live in. You will wonder what semantic type to use for the column containing the counties when you define the data model in Talend Data Stewardship. You want here to add a semantic type specific to your data: `UK_counties` semantic type in this case.

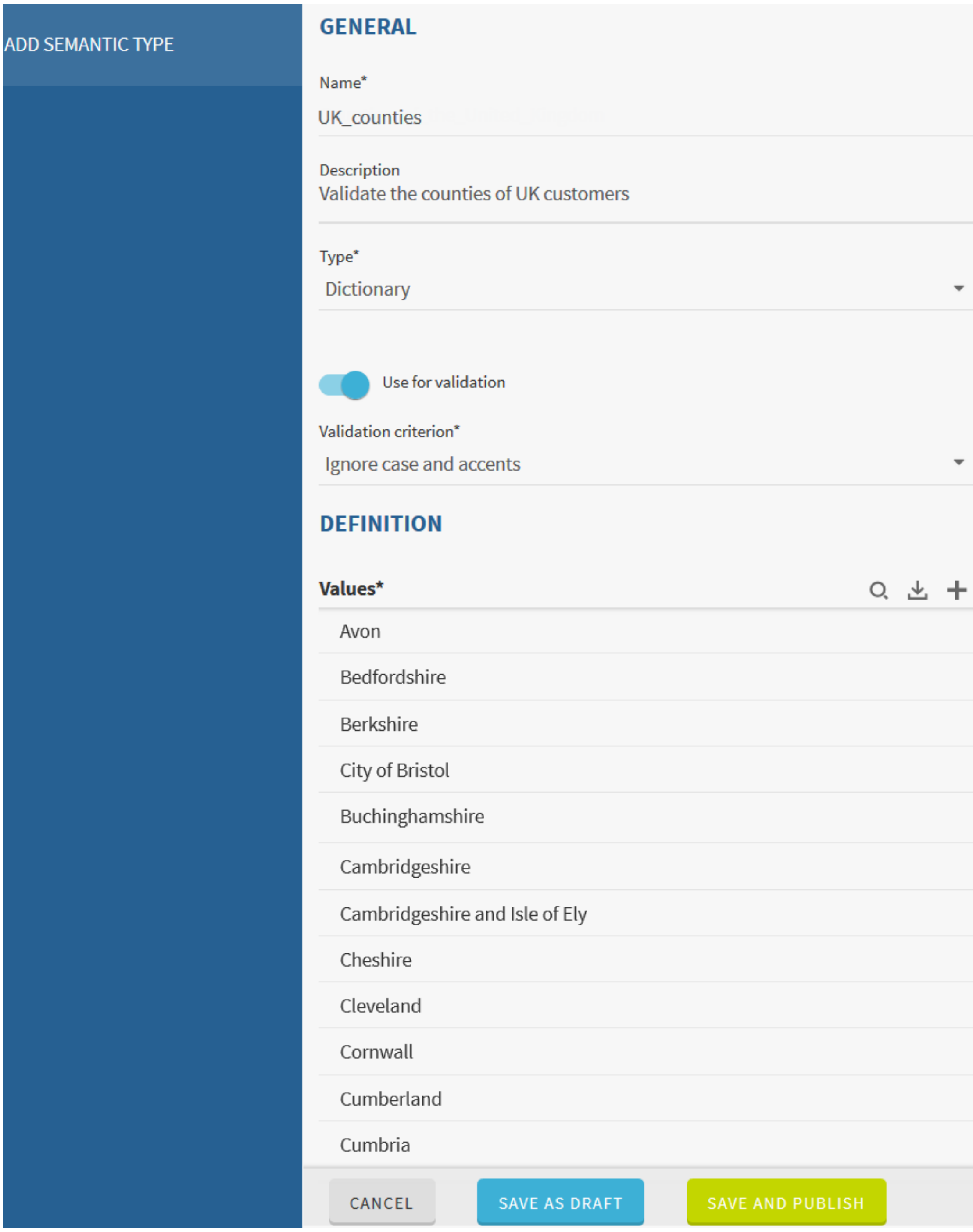
You will create this new semantic type in Talend Dictionary Service, and it will be automatically available in Talend Data Stewardship so that your data can be matched with and validated against a proper type.

1. Create a text file where you list the counties of United Kingdom.

   The file can have one or multiple values per line. Maximum length for a value is 255 characters.

   When you use multiple values on the same line, separate them by commas. In that case, all values are considered as synonyms. You should include in quotes non-alphabetical values, otherwise the file will be rejected.

2. In the home page, click **SEMANTIC TYPES** > **ADD SEMANTIC TYPES**.



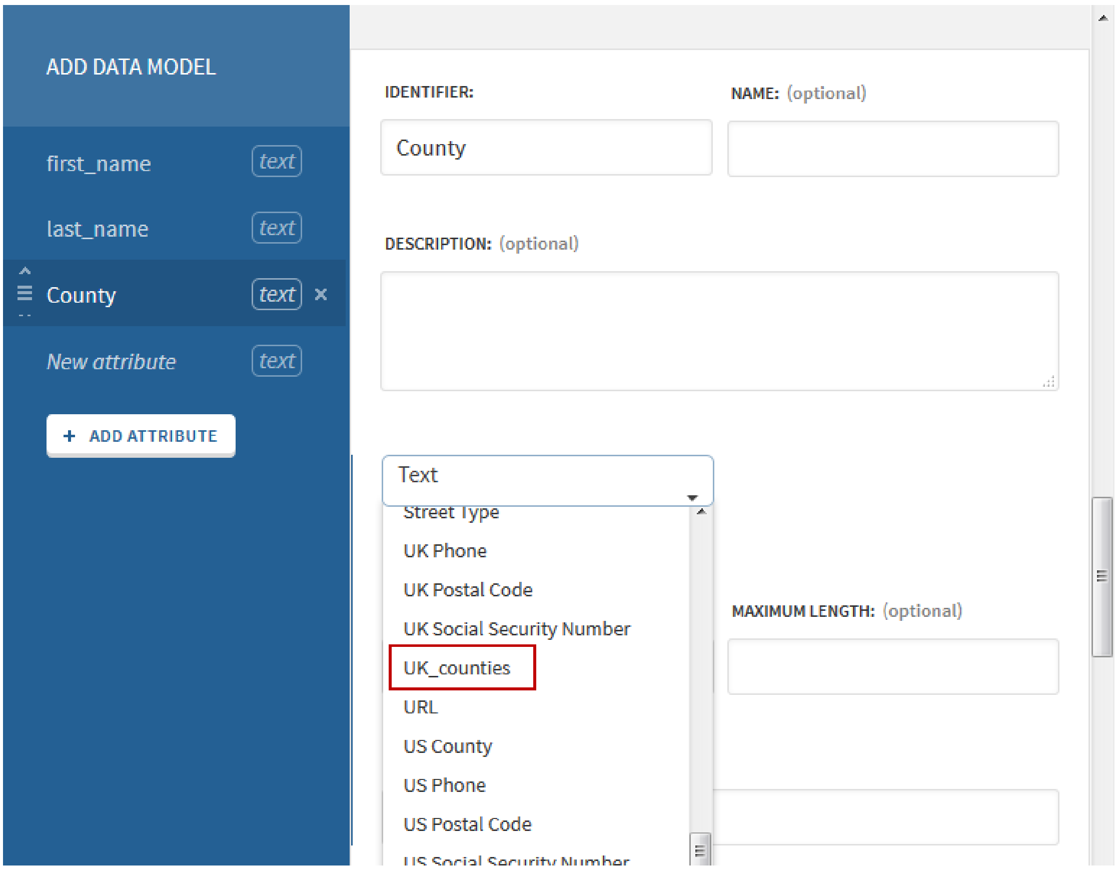
3. Enter a name and a description for the new semantic type.
4. Select the semantic type from the **Type** list.
5. Leave the switch next to **Use for validation** activated.
   Talend Data Stewardship uses semantic types only for validation.
6. From the **Validation criterion** list, select the rule to use while matching data against the values in the dictionary:

Option	Description
Simplified text	Punctuation, white spaces, case and accent are ignored during validation and data is considered as valid. For instance, if `Pâté-en-croûte` is the reference value in the dictionary, then `pate-en-croute` and `PATE--EN CROUTE`

Option	Description
	will both be considered valid but `Pâté n croûte` will not be considered valid.
Ignore case and accents	Case and accents are ignored during validation and data is considered as valid. For instance, if `Pâté-en-croûte` is the reference value in the dictionary, then `pate-en-croute` is considered valid (despite the case and accent differences), but `pate en croute` is not because the dashes have been replaced with spaces.
Exact value	Very restrictive. Data is considered as valid only if it is an exact match with the value.

7. Click the icon to the right of **Values** and import the text file of the counties of United Kingdom.

   You can use the icon to add values manually and the search icon to search values in the list.
8. Click **SAVE AND PUBLISH** to send the semantic type to the Talend Dictionary Service server and make it available to be used by the system.
   Clicking **SAVE AS DRAFT** stores the new type on the server without propagating it to the system. The new type is not usable unless it is published. For a use case of this option, let's say that you have new semantic types to deploy as part of a new project. You can prepare the work by creating the semantic types and save them as draft before the go-live of the project, and can deploy the semantic types only the day of go-live.
9. Go back to Talend Data Stewardship and create a data model for the United Kingdom customers data.



`UK_counties` is now available in the list of the semantic types and you can set it for the `County` column.

When you load data containing the United Kingdom counties to Talend Data Stewardship, they are matched with and validated against the proper semantic type, `UK_counties` that you manually created in Talend Dictionary Service.

Adding values to a dictionary-based semantic type

You can add values to a dictionary-based semantic type you created on the Talend Dictionary Service server either by uploading a file which holds the new values or by entering the values manually. The first alternative saves you much time if you have many values to add to the dictionary-based semantic type. However, duplicate values are not allowed in this semantic type as they are useless and can slow down the process.

1. Create a text file to list the values you want to add to the dictionary-based semantic type you already created.

   The file can have one or multiple values per line. Maximum length for a value is 255 characters.

   When you use multiple values on the same line, separate them by commas. In that case, all values are considered as synonyms. You should include in quotes non-alphabetical values, otherwise the file will be rejected.
2. In the home page, click **SEMANTIC TYPES**.
3. Use the filter field on the top right corner to search for the semantic type you want to modify and click it.
4. Click the icon to the right of **Values** and import the file which holds the extra values.

   You are prompted to append the values contained in the file to the existing ones or replace existing values with the file content.

   Duplicate values are not allowed. When importing the file, a deduplication process is performed automatically and a message is displayed with the number of values added to the dictionary and the number of duplicates found.

   If you have only a few values to add, you can use the + icon and add them manually on the condition that they are not duplicated.
5. Click **Import Values** in the message that opens to confirm the operation.
6. Click **SAVE AND PUBLISH** to send the updated semantic type to the Talend Dictionary Service server and make it available to be used by the system.
   Clicking **SAVE AS DRAFT** stores the new type on the server without propagating it to the system. The new type is not usable unless it is published..

All new values are added to the list open in the user interface.

Adding a new compound semantic type

You can create a compound semantic type which references other semantic types that are published on the Talend Dictionary Service server and add it to the list of recognized data types in the data models in Talend Data Stewardship.

You can mix all semantic types when creating a compound type, and a compound semantic type can reference other compound types on the condition that all children types are already published.

Let's say that you have a file which holds information about customers from US, UK, Germany and France. You need to intervene and validate the different zip codes against a compound semantic type you create. Once data matches one of the child types, it is considered as valid and it is not evaluated against the other referenced types.

When defining the data model in Talend Data Stewardship, you can set the semantic type for the column containing the zip codes to this new compound type, `Zip_codes` in this example.

- All the children semantic types you want to use in the compound type are created and published.

1. In the home page, click **SEMANTIC TYPES** > **ADD SEMANTIC TYPES**.

2. Enter a name and a description for the new semantic type.
3. Select the semantic type from the **Type** list.
4. Leave the switch next to **Use for validation** activated.
   Talend Data Stewardship uses semantic types only for validation.
5. From the **Children types** list, select the semantic types you want to group in this compound type.
6. Click **SAVE AND PUBLISH** to send the semantic type to the Talend Dictionary Service server and make it available to be used by the system.
   Clicking **SAVE AS DRAFT** stores the new type on the server without propagating it to the system. The new type is not usable unless it is published. For a use case of this option, let's say that you have new semantic types to deploy as part of a new project. You can prepare the work by creating the semantic types and save them as draft before the go-live of the project, and can deploy the semantic types only the day of go-live.
7. Go back to Talend Data Stewardship and create the data model for the customers data.
   The new semantic category `Phone_numbers` is available now in the list of semantic types and you can set it for the column containing the phone numbers.

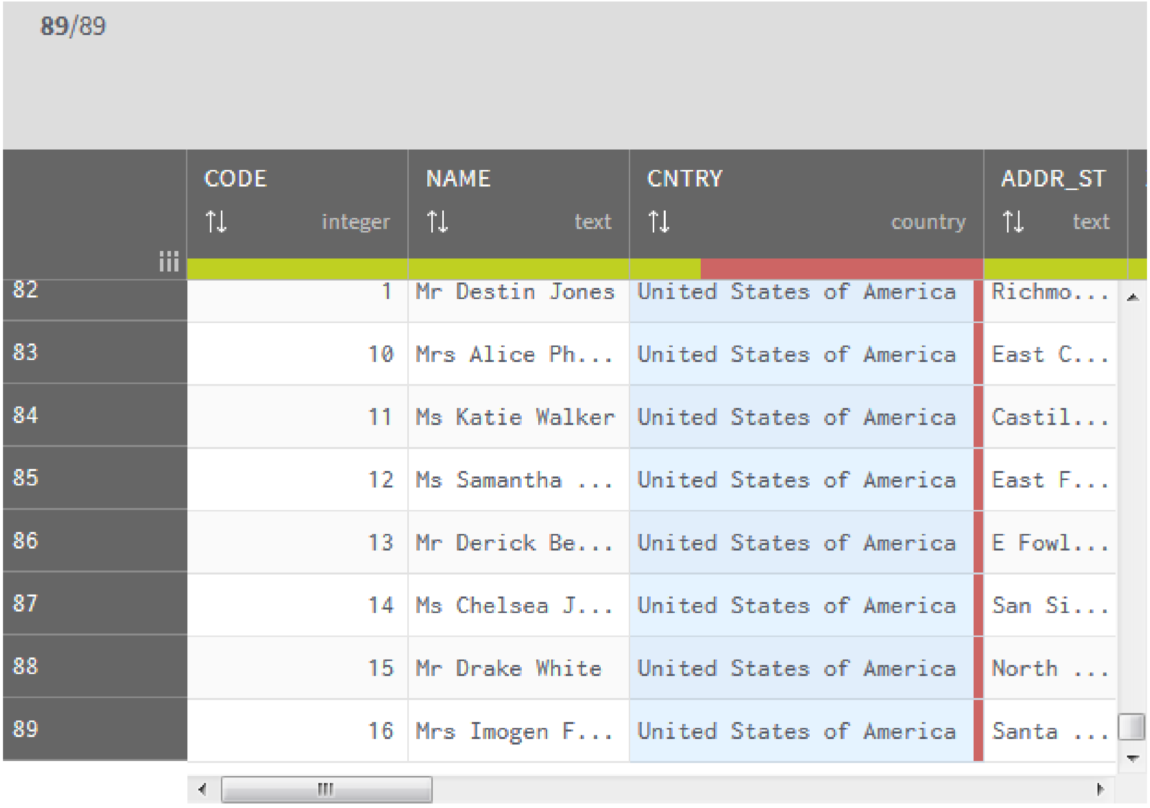
When you load the customer data to Talend Data Stewardship, data is matched and validated against the `Phone_numbers` compound type you created. Data is evaluated against the first child type and if data matches it is not evaluated against the other referenced types and so on.

Updating an existing semantic type

You can edit an existing semantic type in Talend Dictionary Service to impact how your data is validated in Talend Data Stewardship.

Predefined semantic types in Talend Data Stewardship are based on standard values, but you may need to tailor them to match your own data. Some data that you would expect to fall under a predefined category, may be considered invalid.

Let's take the example of a dataset containing a list of customers, with their email addresses, date of birth, and the country they live in. You can notice that all the entries for **United States of America** are considered invalid, when they should not since it is the official name of the country.

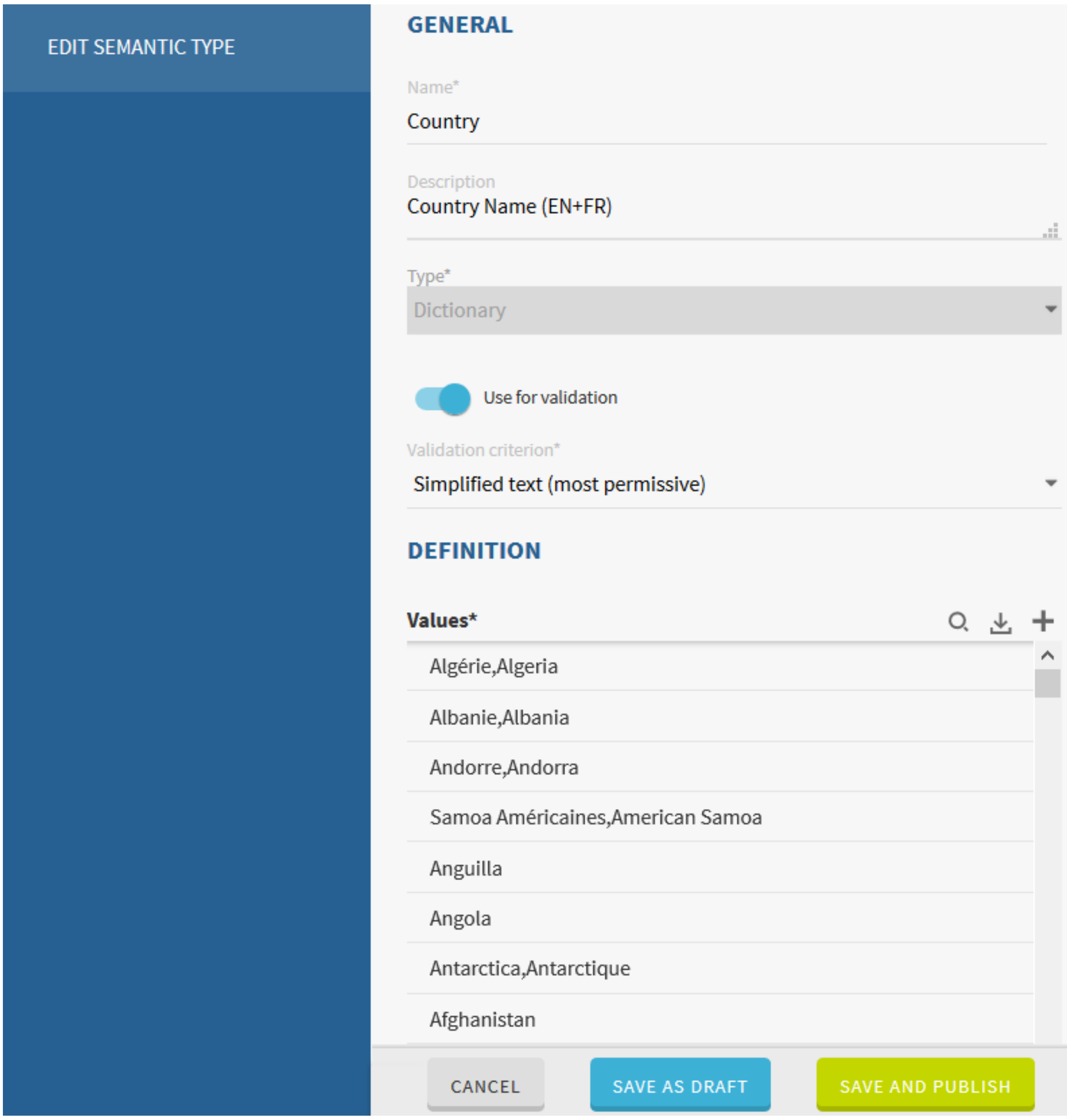

89/89

	CODE integer	NAME text	CNTRY country	ADDR_ST text
82	1	Mr Destin Jones	United States of America	Richmo...
83	10	Mrs Alice Ph...	United States of America	East C...
84	11	Ms Katie Walker	United States of America	Castil...
85	12	Ms Samantha ...	United States of America	East F...
86	13	Mr Derick Be...	United States of America	E Fowl...
87	14	Ms Chelsea J...	United States of America	San Si...
88	15	Mr Drake White	United States of America	North ...
89	16	Mrs Imogen F...	United States of America	Santa ...

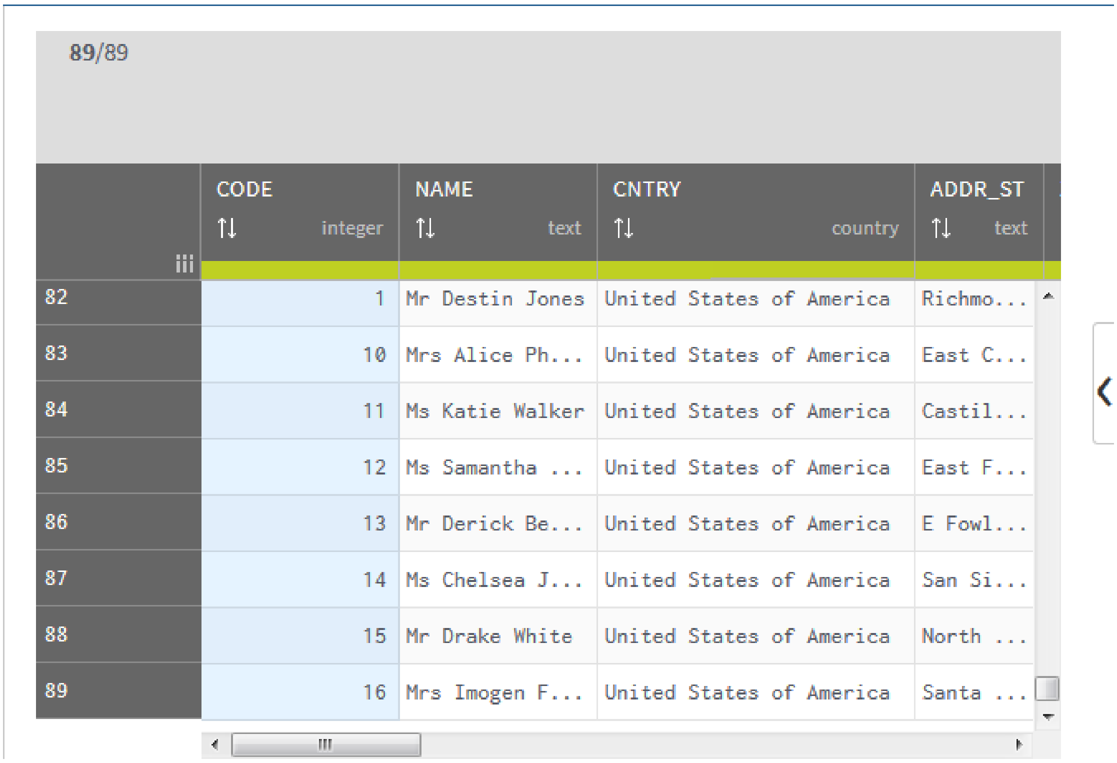
The problem here is that **United States of America** is not one of the expected value for the `Country` semantic type in Talend Data Stewardship. The valid entry in this case would be **United States**.

To avoid having this problem in the future, you need to update the `Country` semantic type in Talend Dictionary Service and add `United States of America` to the list of valid entries. The change will be automatically available in Talend Data Stewardship.

1. In the home page, click **SEMANTIC TYPES**.
2. Click the search icon on the top-right corner of the page and enter `country` to filter the list of semantic types.
3. Click **Country** in the list.



4. Click the icon next to **Values** and enter `United States of America` in the field which displays.
5. Click to add the new value to the top of the list of valid entries for the `Country` semantic type.
6. Click **SAVE AND PUBLISH** to send the semantic type to the Talend Dictionary Service server and make it available to be used by the system.
   Clicking **SAVE AS DRAFT** stores the new type on the server without propagating it to the system. The new type is not usable unless it is published. For a use case of this option, let's say that you have new semantic types to deploy as part of a new project. You can prepare the work by creating the semantic types and save them as draft before the go-live of the project, and can deploy the semantic types only the day of go-live.
7. Go back to Talend Data Stewardship and refresh the task list containing the customers countries or reopen it.



The change in the semantic type is now available in Talend Data Stewardship and you can see in the quality bar that there is no invalid values anymore.

The `Country` semantic type has been manually updated to support the new value.

From now on, when working with data that are matched with the `Country` semantic type, **United States of America** will be considered a valid value.

Removing a semantic type

You can delete a semantic type in Talend Dictionary Service to remove it from the list of recognized data types in Talend Data Stewardship.

You can delete both predefined semantic types as well as predefined standard types.

The variety of semantic types that are present by default in Talend Data Stewardship can be troublesome in certain situations. For example, a five-digit number can be interpreted as a American ZIP code, but also as a French or German one since they share the same format.

Let's say that you are working in an American company, and you only have to deal with data coming from American clients, including ZIP codes. You would prefer to keep only the American ZIP code in the list of recognized semantic types.

Using Talend Dictionary Service, you will simply remove the other semantic types that match the five-digit format and only leave `US Postal Code`. The change will then be ported instantly in Talend Data Stewardship, and you will from now on validate a ZIP code column against the semantic type `US Postal Code`.

1. In the home page, click **SEMANTIC TYPES**.
2. Click the search icon on the top-right corner of the page and enter `postal` to filter the list of semantic types and display only the Zip code types.
3. Point to the name of a Zip code semantic type and click the trash can which displays to the right.

4. Accept to remove the semantic type when prompted and repeat the operation to remove the other semantic types that match five-digit numbers till you are left with only `US Postal Code`.

   When you delete a semantic type which is already used in a data model attached to a campaing, the attributes in the data model automatically switch to `text` and the validity of all tasks is recomputed. This means data which could display as invalid with the initial semantic type may look as valid with the `text` semantic type.

You have deleted all the semantic types compatibles with five-digit numbers but one. From now on, when adding new data models, you can set only `US Postal Code` as the semantic type for columns with Zip code data.

Predefined standard and semantic types

The data models in Talend Data Stewardship define the structure of the data you load to the campaigns and decide the type of the data based on some supported standard and semantic types.

The table below lists all the standard and semantic types supported and integrated by default in Talend Data Stewardship.

Semantic type	Description
Airport Code	Airport code
Amex Card	American Express card
AT VAT Number	Austrian VAT number
Bank Routing Transit Number	Bank routing transit number
BE Postal Code	Belgian postal code
BG VAT Number	Bulgarian VAT number
Boolean	Answers with the value `True` or `False`e
CA Province Territory	Canadian province
CA Province Territory Code	Canadian province code
Civility	Civility
Color Hex Code	Color hexadecimal code
Continent	Continent name
Continent Code	Continent code
Country	Country name
Country Code ISO2	2-letter country code
Country Code ISO3	3-letter country code
Currency Code	Currency code

Semantic type	Description
Currency Name	Currency name
Data URL	URL starting with the word data
Date	Date including day, month and year
Decimal	Decimal numeric value
DE Postal Code	German postal code
De Phone	German phone number
Email	Email address
EN Month	Month in English
EN Month Abbrev	English month abbreviation
EN Weekday	Week day or their abbreviation
File URL	File URL
FR Postal Code	French postal code
FR VAT Number	French VAT number
FR Phone	French phone number
FR Insee Code	French Insee code of cities with Corsica and colonies
FR Commune	French municipality
FR Department	French department
FR Region	French region
FR Region Legacy	Former French regions, prior to the 2016 territorial reform.
FR Social Security Number	French social security number
Gender	Gender
Geographic Coordinate	Geographic coordinate, longitude or latitude coordinates with at least meter precision
Geographic Coordinates	Geographic coordinates, Google Maps style GPS Decimal format

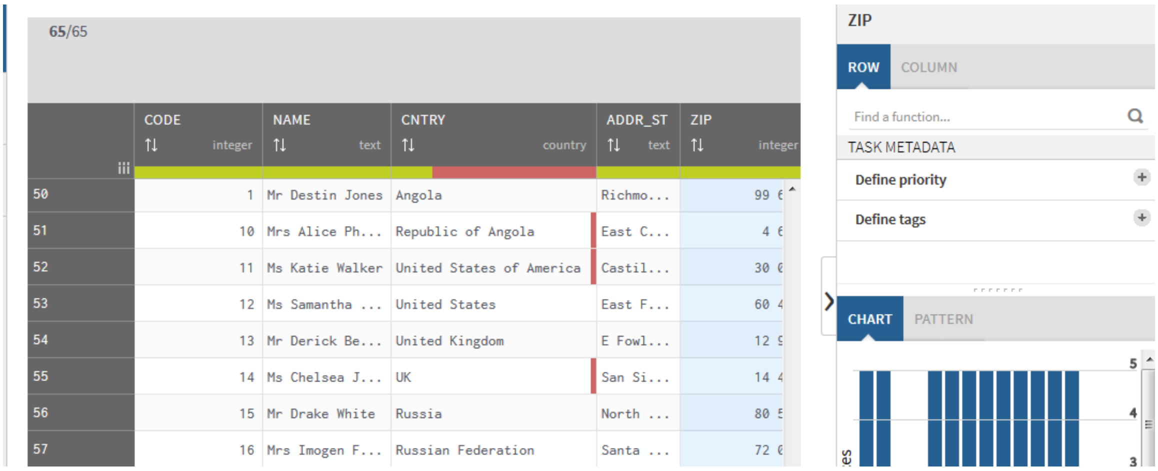
Semantic type	Description
Geographic Coordinates (degree)	Geographic coordinates (degrees), Latitude and longitude coordinates separated by a comma in the form: N 0:59:59.99,E 0:59:59.99"
HDFS URL	HDFS URL
HR_DEPARTMENT	HR department
Integer	Numeric value
IBAN	International Bank Account Number
IPv4 Address	IPv4 address
IPv6 Address	IPv6 address
ISBN-10	International standard book number 10 digits
ISBN-13	International standard book number 13 digits
Language	Language
Language Code ISO2	2-letter language code
Language Code ISO3	3-letter language code
MailTo URL	MailTo URL
Money Amount (EN)	Amount of money in English format
Money Amount (FR)	Amount of money in French format
MAC Address	MAC address
MasterCard	Mastercard credit card
Month	Month
MX Estado	Mexican state
MX Estado Code	Mexican state code
North American state	US state and Canadian province
North American state code	US state code and Canadian province code
Passport	Passport number
Phone number	Phone number (DE, FR, UK, US)
SE Social Security Number	Swedish person number

Semantic type	Description
SEDOL	Stock exchange daily official list
Street Type	Street type
Text	String text
Time	Time of the day
Timestamp	Date and time
UK Postal Code	UK postal code
UK Phone	UK phone number
UK Social Security Number	National identification number, national identity number, or national insurance number generally called NI number
US Postal Code	US postal code
US Phone	US phone number
US Social Security Number	US social security number
URL	Web site URL
US County	US county name
US State	US states
US State Code	US state code
Visa Card	Visa credit card
Web Domain	Web site domain
Weekday	Day of the week

Enriching the semantic types libraries through the command line interface

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.

When campaign owners define the structure of the data to be managed in a campaign, they can select from a predefined list the semantic type for each attribute. Then when they load data into Talend Data Stewardship, an internal validation of the schema type is performed and data is displayed as valid or invalid accordingly.



For example, the list of entries included by default in Talend Data Stewardship under countries does not include `Republic of Angola`, `United States of America` and `UK`. As a result, such entries are considered invalid country names when loaded to Talend Data Stewardship. For further information about the semantic types present by default, see Predefined standard and semantic types on page 34.

But, you can go further and create your own semantic types, as well as updating or deleting the existing ones, so that your experience with Talend Data Stewardship speaks your business language. You can do all these management options either through a user interface integrated in Talend Data Stewardship or through the command line interface.

For further information about using the user interface, see Enriching the semantic types libraries through the user interface on page 23.

Talend Dictionary Service stores all the semantic libraries used in various Talend products, including Talend Data Stewardship. All the changes that you make on the Talend Dictionary Service server are instantly available in Talend Data Stewardship. The availability of Talend Dictionary Service depends on the license you have.

On the server, the semantic types are divided into several categories:

- The `Dictionary` type, based on a closed list of values.
- The `Regular expression` type that compares your data against a preselected regular expression.
- The `Compound type` that compares your data against several semantic types referenced in the compound type.

To display a list of all the available commands in Talend Dictionary Service, go to <Dictionary_Service_Path>/command-line and enter the following command according to your operating system:

- `category_manager.bat -h` command for Windows.
- `./category_manager.sh -h` for Linux.

To enable the interaction between Talend Dictionary Service and Talend Data Stewardship, you must fulfill the following prerequisites:

- Talend Dictionary Service is installed and running.
- Talend Administration Center is installed and running.
- Your license in Talend Administration Center includes Data Quality.

- The role assigned to you in Talend Administration Center is either **Designer** or **Operation manager**, and your user type is **Data management** or above.
- The **Data Stewardship User** check box is selected for your user in the administration center with any of the two possible roles set in the **Data Stewardship Role** field.
- In the <install_folder>\tds\apache-tomcat\conf\data-stewardship.properties file, the dataquality.dictionaryservice.enable property is set as true.

Adding a new regular expression-based semantic type

You can create a semantic type based on a regular expression in Talend Dictionary Service and add it to the list of recognized data types in Talend Data Stewardship.

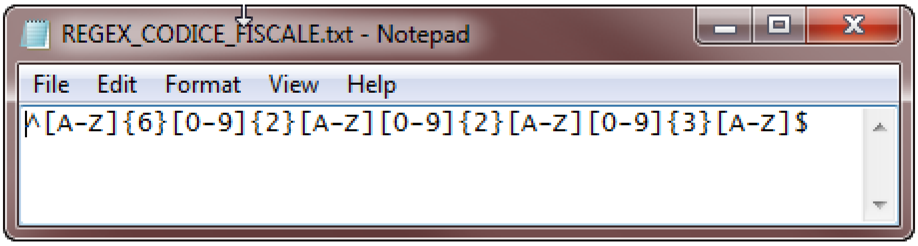
In Talend Dictionary Service, not every type of data can currently be matched with and validated against one of the predefined semantic types. Italian social security numbers, also known as codice fiscale, are currently not recognized for example.

Let's say that you work for an Italian company, only dealing with Italian customers. In this example, you need to intervene and manage some customer data, such as their names, email address, or their social security number.

When defining the data model in Talend Data Stewardship, you will be obliged to set the semantic type for the column containing the social security number to text as there is no predefined semantic type for Italian social security number. This is a bit disappointing and you would like to create a more specific category in order to match this type of data: a codice_fiscale semantic type in this case.

You will create this new semantic type in Talend Dictionary Service, and it will be automatically available in Talend Data Stewardship so that your data can be matched with and validated against a proper type.

1. Create a .txt file containing the following regular expression and save it as REGEX_CODICE_FISCALE.txt.

   This regular expression is designed to match the Italian codice fiscale, which is an alphanumeric code of 16 characters.
2. Add this file to the <Dictionary_Service_Path>/command-line/samples/source folder.

   This folder is used for the sake of this example, but you can save it to your preferred location.
3. Open a command prompt window and use the cd command to go to the <Dictionary_Service_Path>/command-line folder.
4. To create the new codice_fiscale semantic type in Talend Dictionary Service and configure its different parameters, put the following command in one single line and execute it according to your operating system:

   - category_manager.bat -c -name codice_fiscale -type REGEX -desc "Italian social security number" -src samples\source \REGEX_codice_fiscale.txt for Windows.

- `./category_manager.sh -c -name codice_fiscale -type REGEX -desc "Italian social security number" -src samples/source/REGEX_codice_fiscale.txt` for Linux.

  You are prompted for your Talend Administration Center credentials. The command is executed after you enter a valid login and password.

  The `codice_fiscale` semantic type is now added to the list of categories in Talend Dictionary Service.

5. Go back to Talend Data Stewardship and create the data model for the Italian customers data. The new semantic category `codice_fiscale` is available now in the list of semantic types and you can set it for the column containing the social security number.

When you load the customer data to Talend Data Stewardship, data is now matched with and validated against the `codice_fiscale` semantic type, that you created in Talend Dictionary Service.

To display a list of all the available commands in Talend Dictionary Service, go to `<Dictionary_Service_Path>/command-line` and enter the following command according to your operating system:

- `category_manager.bat -h` command for Windows.
- `./category_manager.sh -h` for Linux.

Adding a new dictionary-based semantic type

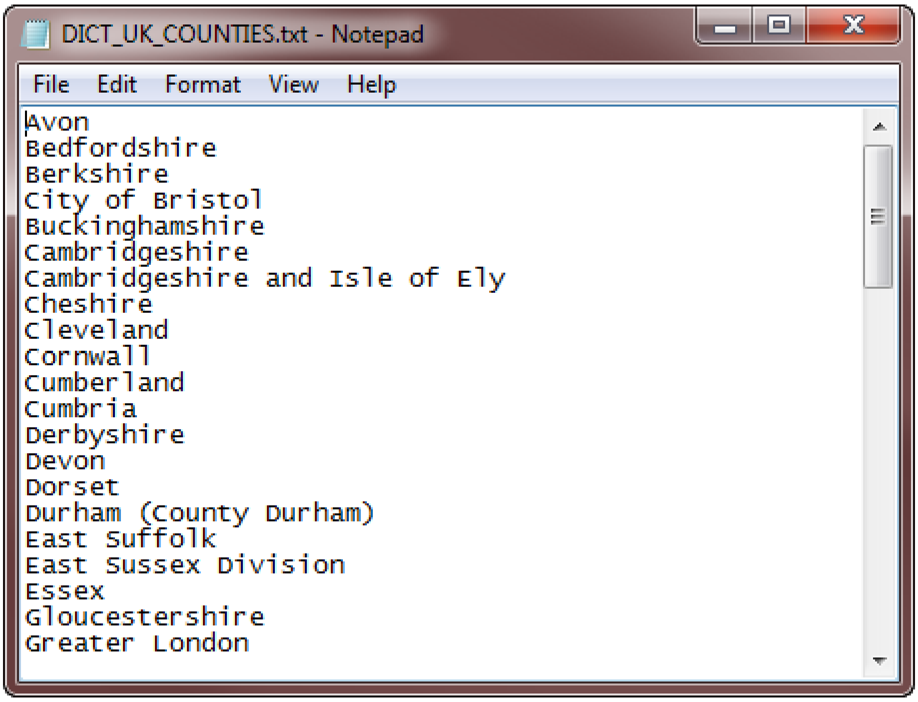
You can create a semantic type based on a dictionary in Talend Dictionary Service and add it to the list of recognized data types in Talend Data Stewardship. However, duplicate values are not allowed in a dictionary-based semantic type as they are useless and can slow down the process.

In Talend Data Stewardship, not every type of data can currently be matched with one of the predefined semantic types. The counties of United Kingdom for example, are currently not recognized as such.

Let's say that you work for a British company, with customers only residing in the United Kingdom. In this example, you need to intervene and manage some customer data, such as their names, email address, or the county they live in. You will wonder what semantic type to use for the column containing the counties when you define the data model in Talend Data Stewardship. You want here to add a semantic type specific to your data: `UK_counties` semantic type in this case.

You will create this new semantic type in Talend Dictionary Service, and it will be automatically available in Talend Data Stewardship so that your data can be matched with and validated against a proper type.

1. Create a .txt file containing the exhaustive list of the British counties and save it as DICT_UK_COUNTIES.txt.
   Make sure to enter one item per line.

   Unlike an open dictionary which purpose is to identify data, this exhaustive list acts as a closed dictionary of values to validate data in Talend Data Stewardship. Data that exactly matches one of the listed values is categorized as a British county.
2. Add this file to the <Dictionary_Service_Path>/command-line/samples/source folder.

   This folder is used for the sake of this example, but you can save it to your preferred location.
3. Open a command prompt window and use the cd command to go to the <Dictionary_Service_Path>/command-line folder.
4. To create the new UK_counties semantic type in Talend Dictionary Service and configure its different parameters, put the following command in one single line and execute it according to your operating system:

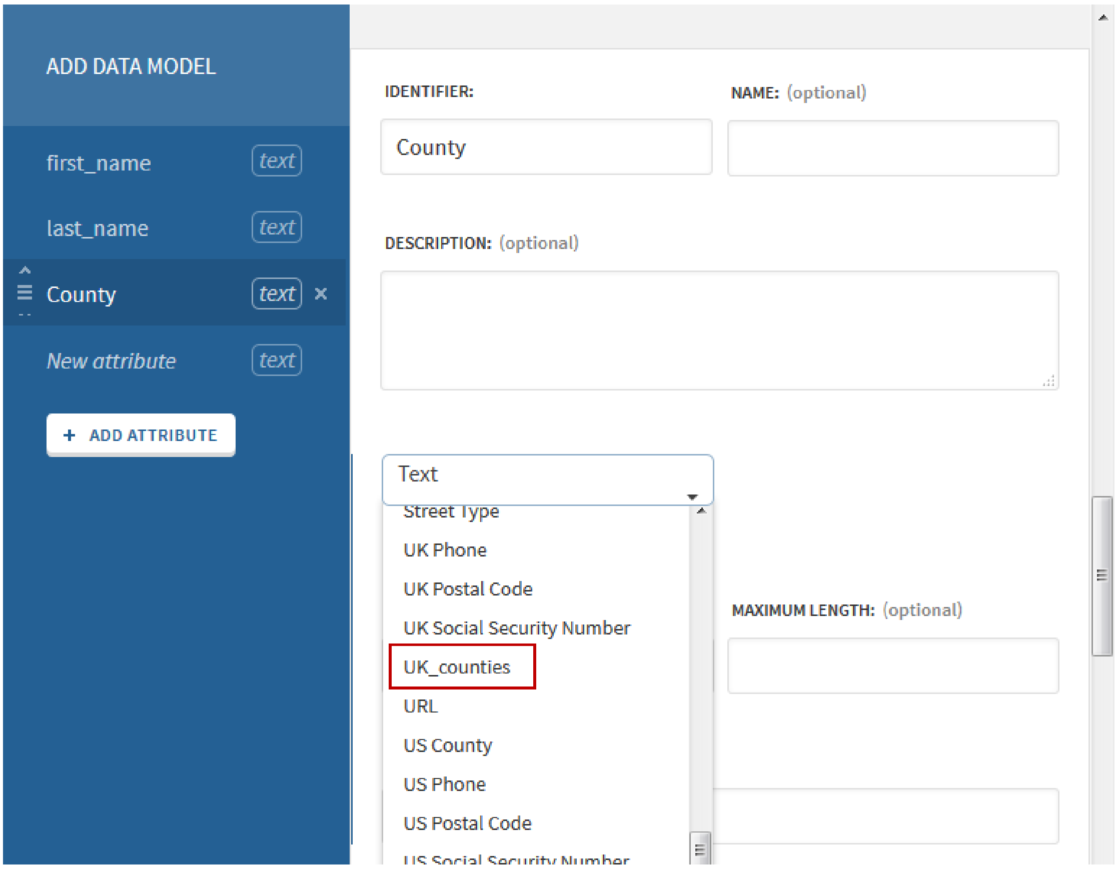
   - category_manager.bat -c -name UK_counties -type DICT -cmpl true -desc "Counties of the United Kingdom" -src samples\source \DICT_UK_COUNTIES.txt for Windows.
   - ./category_manager.sh -c -name UK_counties -type DICT -cmpl true -desc "Counties of the United Kingdom" -src samples/source/ DICT_UK_COUNTIES.txt for Linux.

   You are prompted for your Talend Administration Center credentials. The command is executed after you enter a valid login and password.

   The -cmpl attribute stands for completeness, and is used to determine if the dictionary you are adding is a closed dictionary. It is set to false by default but in this case, it must be set to true. Open dictionaries are not supported with Talend Data Stewardship.

The `UK_counties` semantic type is now added to the list of categories in Talend Dictionary Service.

5. Go back to Talend Data Stewardship and create a data model for the United Kingdom customers data.



`UK_counties` is now available in the list of the semantic types and you can set it for the `County` column.

When you load data containing the United Kingdom counties to Talend Data Stewardship, they are matched with and validated against the proper semantic type, `UK_counties` that you manually created in Talend Dictionary Service.

To display a list of all the available commands in Talend Dictionary Service, go to `<Dictionary_Service_Path>/command-line` and enter the following command according to your operating system:

- `category_manager.bat -h` command for Windows.
- `./category_manager.sh -h` for Linux.

Updating an existing semantic type

You can edit an existing semantic type in Talend Dictionary Service to impact how your data is validated in Talend Data Stewardship.

Predefined semantic types in Talend Data Stewardship are based on standard values, but you may need to tailor them to match your own data. Some data that you would expect to fall under a predefined category, may be considered invalid.

Let's take the example of a dataset containing a list of customers, with their email addresses, date of birth, and the country they live in. You can notice that all the entries for **United States of America** are considered invalid, when they should not since it is the official name of the country.


89/89

	CODE integer	NAME text	CNTRY country	ADDR_ST text
82	1	Mr Destin Jones	United States of America	Richmo...
83	10	Mrs Alice Ph...	United States of America	East C...
84	11	Ms Katie Walker	United States of America	Castil...
85	12	Ms Samantha ...	United States of America	East F...
86	13	Mr Derick Be...	United States of America	E Fowl...
87	14	Ms Chelsea J...	United States of America	San Si...
88	15	Mr Drake White	United States of America	North ...
89	16	Mrs Imogen F...	United States of America	Santa ...

The problem here is that **United States of America** is not one of the expected value for the `country` semantic type in Talend Data Stewardship. The valid entry in this case would be **United States**.

To avoid having this problem in the future, you will update the `country` semantic type in Talend Dictionary Service, and add `United States of America` to the list of valid entries. The change will be automatically available in Talend Data Stewardship.

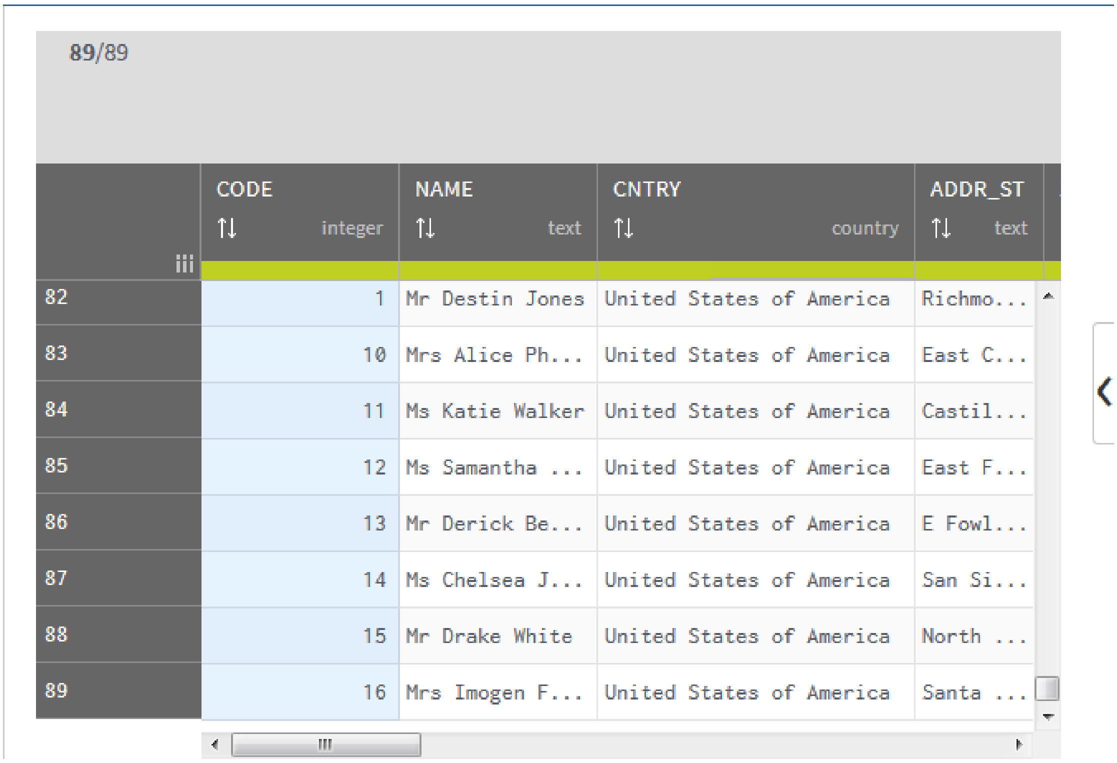
1. Open a command prompt window and use the `cd` command to go to the `<Dictionary_Service_Path>/command-line` folder.
2. To add the value `United States of America` to the list of valid countries, execute the following command according to your operating system:
   - `category_manager.bat -a -name COUNTRY -value "United States of America"` for Windows.
   - `./category_manager.sh -a -name COUNTRY -value "United States of America"` for Linux.

   Please note that to be able to use this command, you need to put it on one single line.

   You are prompted for your Talend Administration Center credentials. The command is executed after you enter a valid login and password.
3. To display the list of entries under the `country` semantic type, execute the following command according to your operating system:
   - `category_manager.bat -e -name COUNTRY` for Windows.
   - `./category_manager.sh -e -name COUNTRY` for Linux.

   You can see that `United States of America` has been properly added at the bottom of the list of valid entries for the `country` semantic type.

4. Go back to Talend Data Stewardship and refresh the task list containing the customers countries or reopen it.



89/89

	CODE integer	NAME text	CNTRY country	ADDR_ST text
82	1	Mr Destin Jones	United States of America	Richmo...
83	10	Mrs Alice Ph...	United States of America	East C...
84	11	Ms Katie Walker	United States of America	Castil...
85	12	Ms Samantha ...	United States of America	East F...
86	13	Mr Derick Be...	United States of America	E Fowl...
87	14	Ms Chelsea J...	United States of America	San Si...
88	15	Mr Drake White	United States of America	North ...
89	16	Mrs Imogen F...	United States of America	Santa ...

The change in the semantic type is now available in Talend Data Stewardship and you can see in the quality bar under the column header that there is no invalid values anymore.

The `country` semantic type has been manually updated to support a new value.

From now on, when dealing with data that are matched with the `country` semantic type, **United States of America** will be considered a valid value.

To display a list of all the available commands in Talend Dictionary Service, go to <Dictionary_Service_Path>/command-line and enter the following command according to your operating system:

- `category_manager.bat -h` command for Windows.
- `./category_manager.sh -h` for Linux.

Removing a semantic type

You can delete a semantic type in Talend Dictionary Service to remove it from the list of recognized data types in Talend Data Stewardship.

You can delete both predefined semantic types as well as predefined standard types.

The variety of semantic types that are present by default in Talend Data Stewardship can be troublesome in certain situations. For example, a five-digit number can be interpreted as a American ZIP code, but also as a French or German one since they share the same format.

Let's say that you are working in an American company, and you only have to deal with data coming from American clients, including ZIP codes. You would prefer to keep only the American ZIP code in the list of recognized semantic types.

Using Talend Dictionary Service, you will simply remove the other semantic types that match the five-digit format and only leave US_POSTAL_CODE. The change will then be ported instantly in Talend Data Stewardship, and you will always from now on validate a ZIP code column against the semantic type US_POSTAL_CODE.

1. Open a command prompt window.
2. Use the cd command, go to the <Dictionary_Service_Path>/command-line folder.
3. To display the names of the existing semantic types and see which ones to remove, execute the following command: according to your operating system:
   - category_manager.bat -l -type REGEX for Windows.
   - ./category_manager.sh -l -type REGEX for Linux.

   You are prompted for your Talend Administration Center credentials. The command is executed after you enter a valid login and password.

   The list of semantic types based on regular expressions is displayed. You can identify the name of the ones you want to remove, FR_POSTAL_CODE or DE_POSTAL_CODE among others.
4. To remove the French postal codes semantic type, execute the following command according to your operating system:
   - category_manager.bat -d -name FR_POSTAL_CODE for Windows.
   - ./category_manager.sh -d -name FR_POSTAL_CODE for Linux.

   The FR_POSTAL_CODE has been removed from the list of recognized semantic types and you can not associate five-digit numbers with French ZIP codes anymore when creating data models in Talend Data Stewardship.
5. Repeat this operation to remove the other semantic types that match five-digit numbers:
   - DE_POSTAL_CODE
   - FR_INSEE_CODE

   When you delete a semantic type which is already used on a column in a data model attached to a camping, the semantic type of the column is automatically set to text. This means data which could display as invalid with the initial semantic type may look as valid with the text semantic type.

You have deleted all the semantic types compatibles with five-digit numbers but one. From now on, when adding new data models, you can set only US_POSTAL_CODE as the semantic type for columns with Zip code data.

To display a list of all the available commands in Talend Dictionary Service, go to <Dictionary_Service_Path>/command-line and enter the following command according to your operating system:

- category_manager.bat -h command for Windows.
- ./category_manager.sh -h for Linux.

Predefined standard and semantic types

The data models in Talend Data Stewardship define the structure of the data you load to the campaigns and decide the type of the data based on some supported standard and semantic types.

The table below lists all the standard and semantic types supported and integrated by default in Talend Data Stewardship.

Semantic type	Description
Airport Code	Airport code
Amex Card	American Express card
AT VAT Number	Austrian VAT number
Bank Routing Transit Number	Bank routing transit number
BE Postal Code	Belgian postal code
BG VAT Number	Bulgarian VAT number
Boolean	Answers with the value `True` or `False`e
CA Province Territory	Canadian province
CA Province Territory Code	Canadian province code
Civility	Civility
Color Hex Code	Color hexadecimal code
Continent	Continent name
Continent Code	Continent code
Country	Country name
Country Code ISO2	2-letter country code
Country Code ISO3	3-letter country code
Currency Code	Currency code
Currency Name	Currency name
Data URL	URL starting with the word data
Date	Date including day, month and year
Decimal	Decimal numeric value
DE Postal Code	German postal code
De Phone	German phone number
Email	Email address
EN Month	Month in English
EN Month Abbrev	English month abbreviation

Semantic type	Description
EN Weekday	Week day or their abbreviation
File URL	File URL
FR Postal Code	French postal code
FR VAT Number	French VAT number
FR Phone	French phone number
FR Insee Code	French Insee code of cities with Corsica and colonies
FR Commune	French municipality
FR Department	French department
FR Region	French region
FR Region Legacy	Former French regions, prior to the 2016 territorial reform.
FR Social Security Number	French social security number
Gender	Gender
Geographic Coordinate	Geographic coordinate, longitude or latitude coordinates with at least meter precision
Geographic Coordinates	Geographic coordinates, Google Maps style GPS Decimal format
Geographic Coordinates (degree)	Geographic coordinates (degrees), Latitude and longitude coordinates separated by a comma in the form: N 0:59:59.99,E 0:59:59.99"
HDFS URL	HDFS URL
HR_DEPARTMENT	HR department
Integer	Numeric value
IBAN	International Bank Account Number
IPv4 Address	IPv4 address
IPv6 Address	IPv6 address
ISBN-10	International standard book number 10 digits
ISBN-13	International standard book number 13 digits

Semantic type	Description
Language	Language
Language Code ISO2	2-letter language code
Language Code ISO3	3-letter language code
MailTo URL	MailTo URL
Money Amount (EN)	Amount of money in English format
Money Amount (FR)	Amount of money in French format
MAC Address	MAC address
MasterCard	Mastercard credit card
Month	Month
MX Estado	Mexican state
MX Estado Code	Mexican state code
North American state	US state and Canadian province
North American state code	US state code and Canadian province code
Passport	Passport number
Phone number	Phone number (DE, FR, UK, US)
SE Social Security Number	Swedish person number
SEDOL	Stock exchange daily official list
Street Type	Street type
Text	String text
Time	Time of the day
Timestamp	Date and time
UK Postal Code	UK postal code
UK Phone	UK phone number
UK Social Security Number	National identification number, national identity number, or national insurance number generally called NI number
US Postal Code	US postal code

Semantic type	Description
US Phone	US phone number
US Social Security Number	US social security number
URL	Web site URL
US County	US county name
US State	US states
US State Code	US state code
Visa Card	Visa credit card
Web Domain	Web site domain
Weekday	Day of the week

Working with tasks

Managing tasks

In Talend Data Stewardship, campaign owners and/or data stewards who are part of a specific campaign can perform some management procedures on the tasks listed in the campaign.

For a list of all available functions on tasks, see List of functions on tasks on page 59.

Transitioning tasks

Tasks have a lifecycle where they pass through different states according to the workflow defined in the campaign. Transitioning defines how a task goes from one state to the next in the workflow. There is always an initial and a final states.

With the two-state workflow, there is only one transition between the initial and the final states, while a three-state workflow requires two transitions between states. With each transition, tasks progress to the next step in the workflow and are made available to the data stewards who are assigned the right to access tasks at that state based on some rules:

- If the target state is not the final Resolved state:
  - if the current participant is allowed to access this state, then the tasks remain assigned to him,
  - if the target state has only one participant, then the tasks remain assigned to this unique participant,
  - otherwise, the tasks are unassigned.
- If the target state is a final state, the tasks are unassigned.

You can perform a transition on a single task or perform the same transition on multiple tasks.

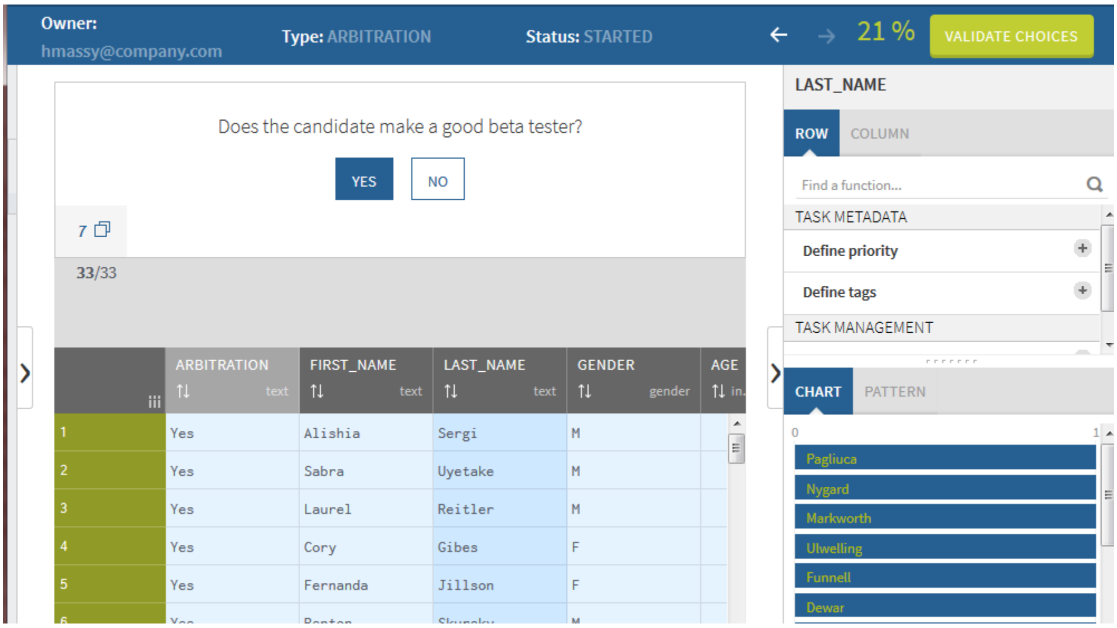
- A campaign owner has created the campaign and granted you access to it.

  For further information, see the online publication about adding campaigns and defining roles on Talend Help Center (https://help.talend.com).

- 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 name of the campaign you want to open.
2. In the task list, select a task or hold the SHIFT key and select multiple tasks if you think you can apply the same transition choice on all of them.
3. Select the transition choice on top of the list. Select YES in this example.
   The transition choice you have depends on at what state you are in the workflow. This example shows bulk transition in an initial state of the workflow in an **Arbitration** campaign.



   The selected tasks are marked as green. You may get a warning message if the transition choice can not be applied to all the selected tasks.
4. Click **VALIDATE CHOICES** on the top-right corner of the page to transition the tasks to the next step in the workflow.

Accessing tasks at a validation state

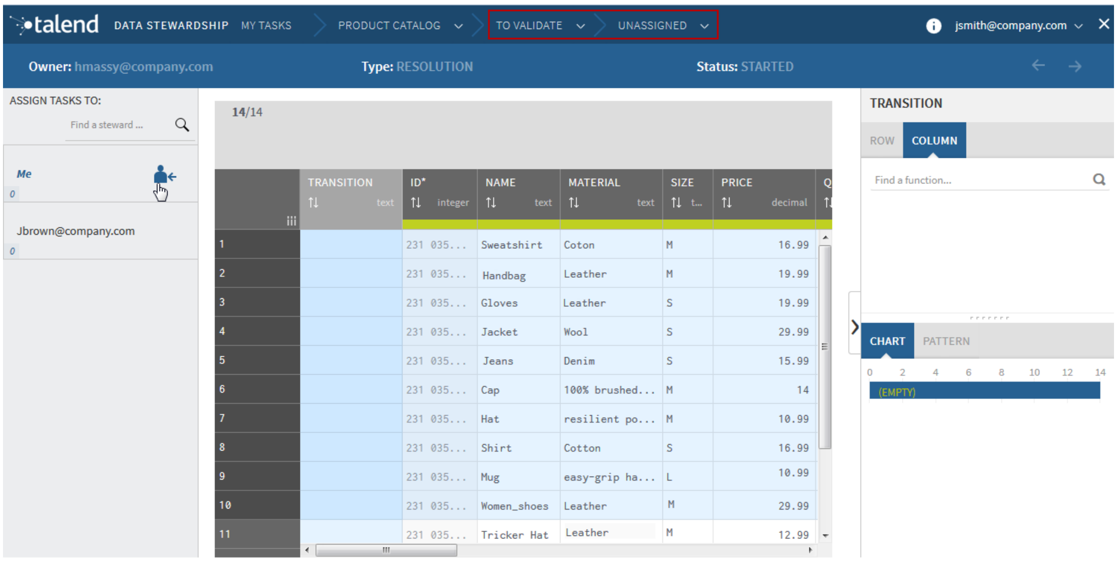
Tasks have a lifecycle where they pass through different states according to the workflow defined in the campaign. In a three-state workflow, only participants who are granted the right role can have access to the intermediate state where tasks need to be validated.

At the validation state, authorized data stewards can perform a transition choice on a single task or on multiple tasks and move them to the final state in the workflow.

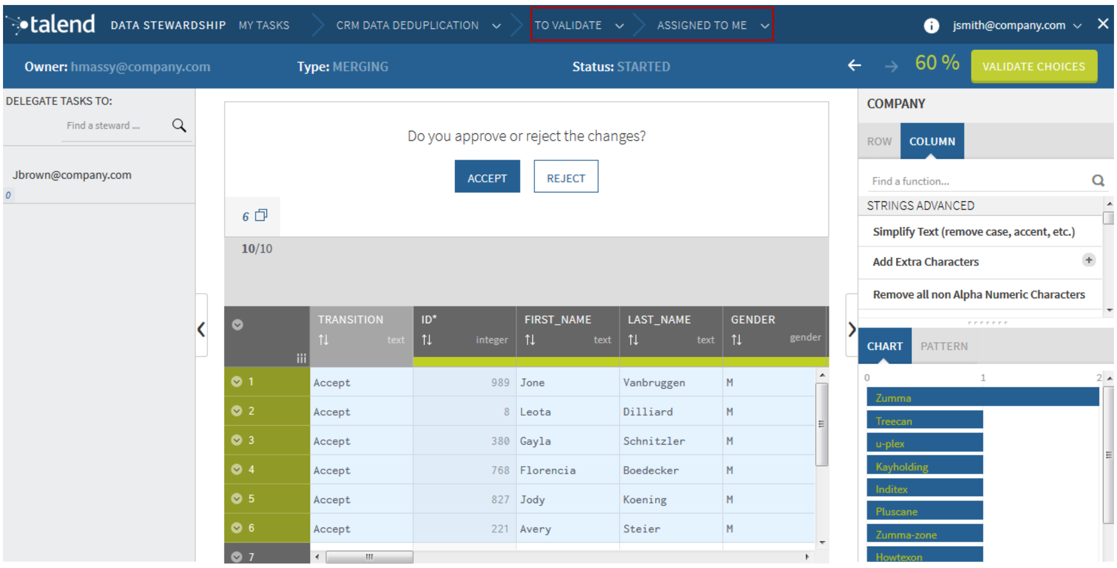
- The workflow defined in the campaign is a three-state workflow, which includes a validation state.
- You are part of the campaign and have been granted the role to access the tasks at the right state in the workflow.

  For further information, see the online publication about defining roles and workflows on Talend Help Center (https://help.talend.com).

1. In the **TASKS** page, click the name of the campaign you want to open.
2. On the menu bar, use the menu combinations **TO VALIDATE** > **UNASSIGNED** to open the list of the tasks which are transitioned to the validation state but are not assigned yet.



3. Select the tasks and click  next to **Me** to move the tasks to your dashboard.
4. On the menu bar, click **UNASSIGNED** and select **Assigned to me** to display the tasks for which you need to validate the choices of the initial data stewards.



5. Select one or more of the tasks and approve the changes and click either **ACCEPT** or **REJECT**. The selected task(s) are marked with green background and the **TRANSITION** column is filled in with the decision.
6. Click **VALIDATE CHOICES** in the top-right corner of the page to validate the transition choice you have done on the tasks.

The approved tasks are moved from the list and transitioned to the final state in the workflow. The rejected tasks are moved back to the initial list of unassigned tasks if there are several data stewards granted the first role in the workflow. However, if only one data steward is granted the first role in the workflow, the rejected tasks are moved back directly to his/her list of new tasks.

Accessing resolved tasks

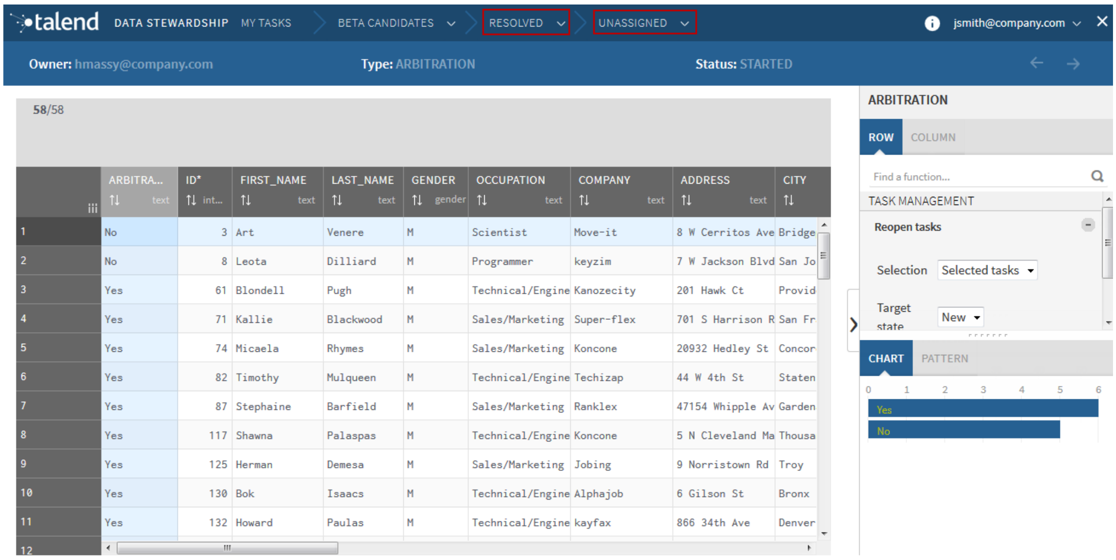
Tasks have a lifecycle where they pass through different states according to the workflow defined in the campaign. When data stewards finalize modifying the tasks and transitioning them through the states defined in the workflow, they reach the resolved state.

Accessing resolved tasks is very important to reopen them once resolved and send them back to the initial state in the workflow.

- You are part of the campaign and have been granted the role to access the tasks at the right state in the workflow.

  For further information, see the online publication about defining roles and workflows on Talend Help Center (https://help.talend.com).

1. In the **TASKS** page, click the name of the campaign you want to open.
2. On the menu bar, click **NEW** and select **Resolved**. Also, click **ASSIGNED TO ME** and select **Unassigned** to open the list of the resolved tasks.



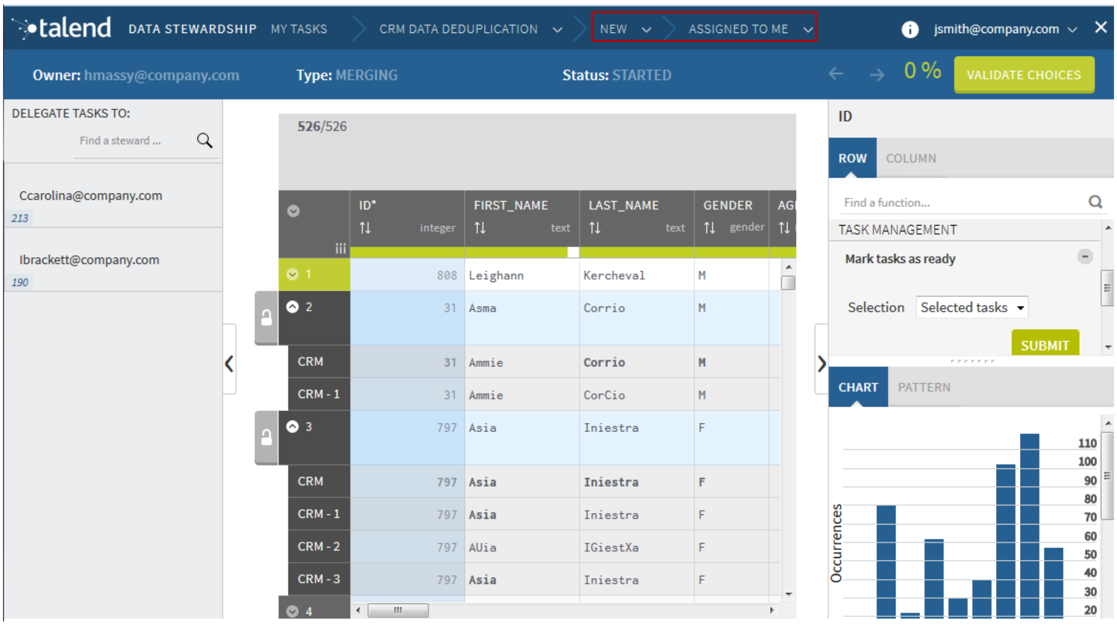
3. Select one or more tasks and click **ROW** in the panel to the right.
4. Reopen the tasks as outlined in Reopening tasks on page 55 .

Marking tasks as ready for validation

When handling tasks in a **Resolution** or **Merging** campaign, data stewards can mark tasks as ready to be validated either individually or all at once on the condition that tasks are in the initial state of the workflow and not in a validation state.

- A campaign owner has defined you as part of a **Resolution** or a **Merging** campaign and has assigned you tasks in the campaign.
- You have accessed your tasks at the initial state of the workflow.

1. In the **TASKS** page, click the name of the campaign you want to open.
2. Depending on the type of the tasks you are handling, fix the values in the fields or build master records from the source data.



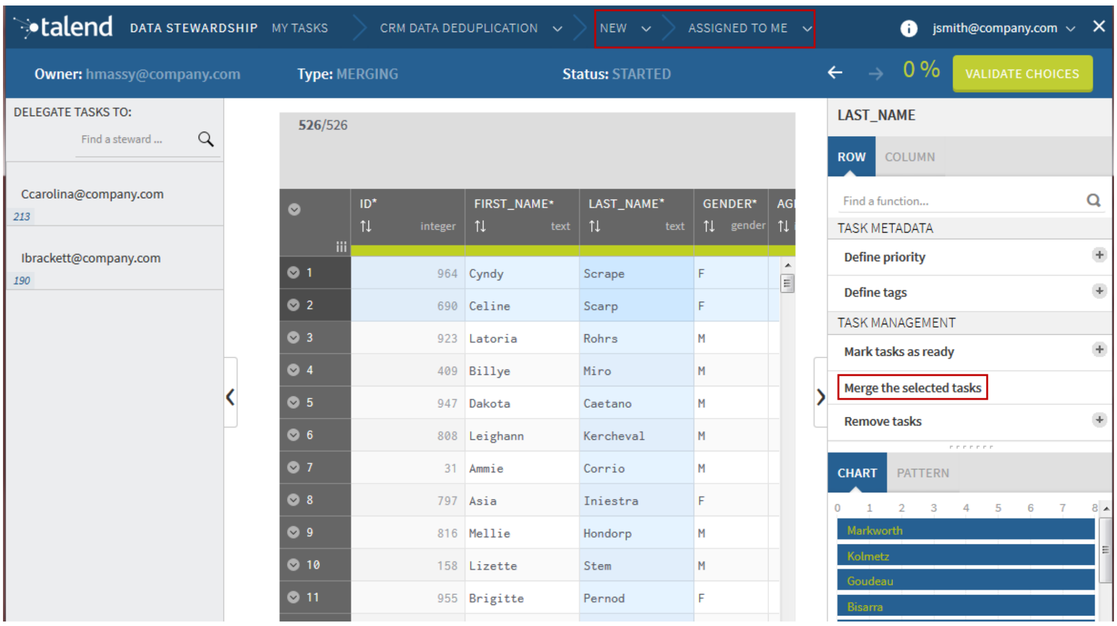
3. In the right panel, click **ROW** and expand **Mark tasks as ready**.
4. From the **Selection** list, select to change the status of all the tasks or only of the tasks you select in the list.
5. Click **SUBMIT** to change the task status and mark them with a green background.
6. If needed, click the **Undo** arrow in the top-right corner to revert the changes you made.

Merging several tasks

When handling tasks in a **Merging** campaign, data stewards may find that the source records of two or more tasks should be merged into a single task.

- A campaign owner has defined you as part of a **Merging** campaign and assigned you some tasks.

1. In the **TASKS** page, click the name of the **Merging** campaign you want to open.
2. In the list of the tasks assigned to you, whether at the initial state or at the validation state of the workflow, select the tasks you want to merge.



3. In the right panel, click **ROW** > **Merge the selected tasks**.
   The values in the new task are selected according to the campaign default survivorship rules applied to all source records. This means that any manual change done in the meantime on the tasks just merged is lost.

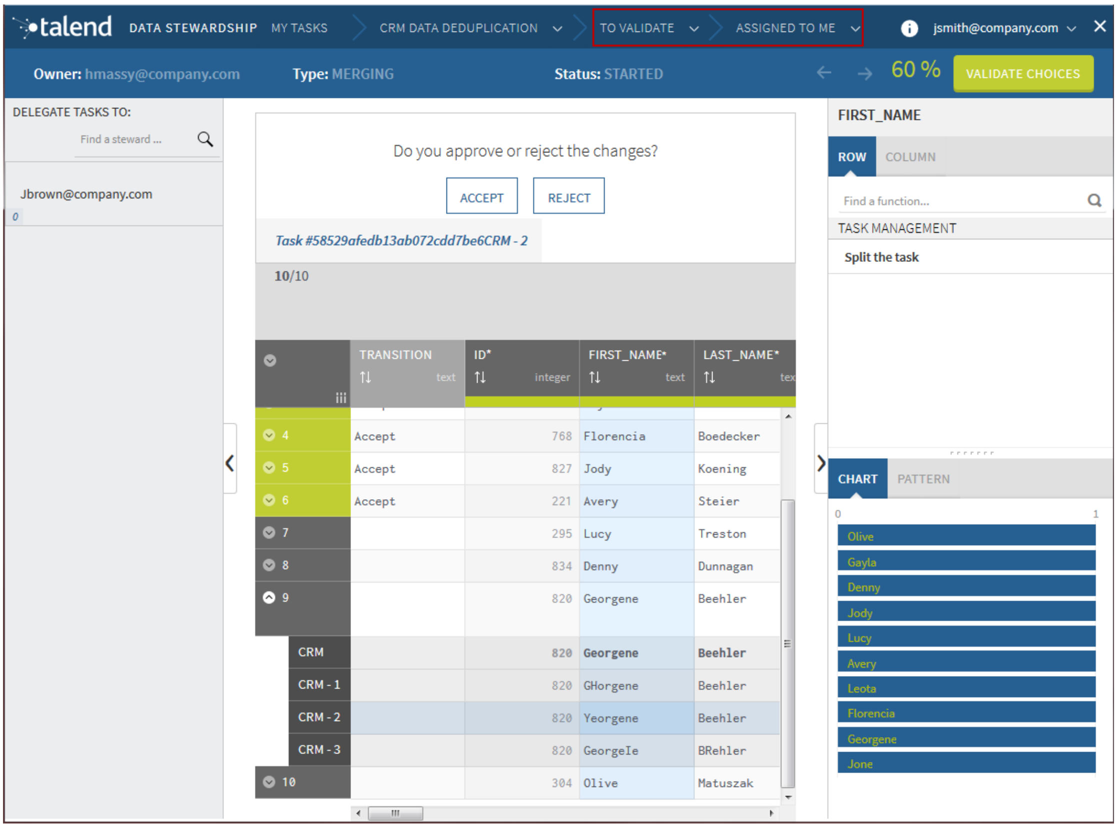
   The selected tasks are merged into one and moved to the bottom of the list.
4. If needed, click the **Undo** arrow in the top-right corner to revert the merge and split the merged task into the initial individual tasks.

Splitting a task

When handling tasks in a **Merging** campaign, data stewards may find that one or more of the source records of a task are inappropriate matches and should belong to a separate task.

- A campaign owner has defined you as part of a **Merging** campaign and assigned you some tasks.

1. In the **TASKS** page, click the name of the **Merging** campaign you want to open.
2. In the list of the tasks assigned to you, whether at the initial state or at the validation state of the workflow, click the drop-down arrow to the left of the task to expand it and display the source records.



3. Select the source record or records you want to move from the current task and group in a new separate task.
4. In the right panel, click **ROW** > **Split the task**.
   The master records for the new task and the initial one are computed as the following:
   - If one task or the other has only one source left, then the new task uses the remaining source values for all attributes,
   - If manual changes were performed on a given attribute, then these manual changes are kept for both the initial and the new tasks,
   - If one source was explicitly selected for a given attribute, then this source value is kept for the task to which it belongs,

- Otherwise, the attribute is computed according to the rules defined in the campaign.

  The selected source records are removed from the current task and grouped in a new task which is added at the bottom of the list.

5. If needed, click the **Undo** arrow in the top-right corner to revert the split and put the source records back under the initial task.

Reopening tasks

Tasks have a lifecycle where they pass through different states according to the workflow defined in the campaign till they reach the resolved state.

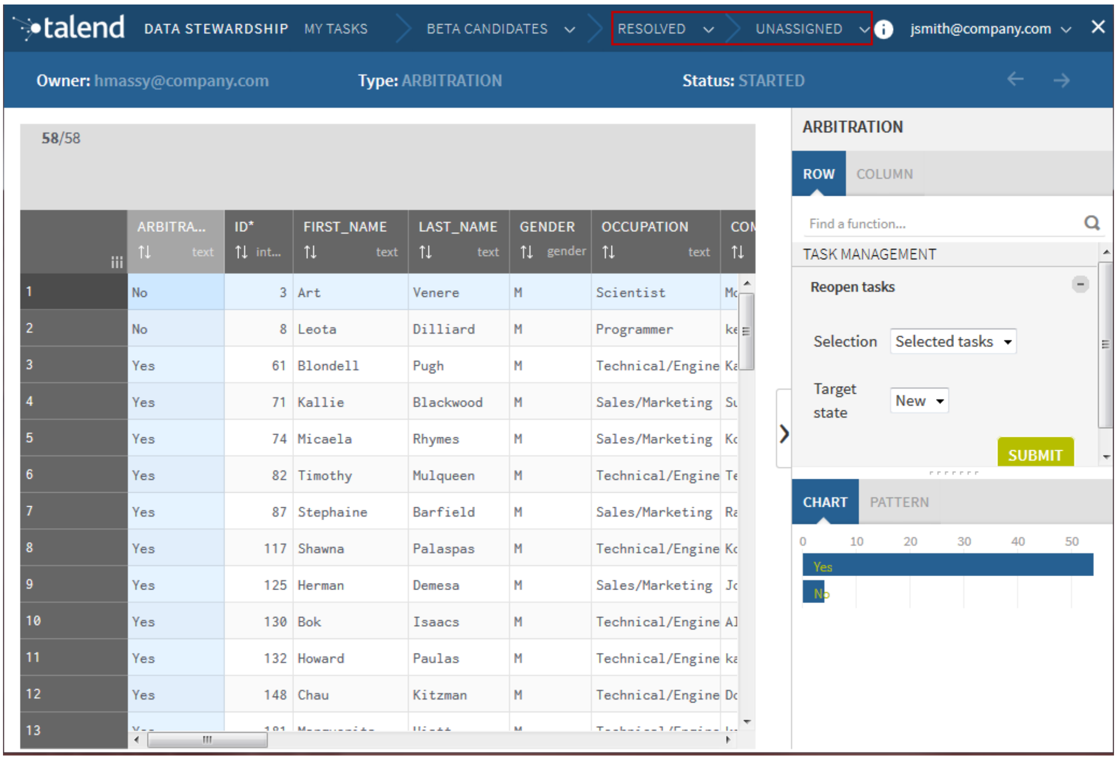
However, campaign owners or data stewards, who are granted the right role, can reopen the tasks once resolved and send them back to a specific state in the workflow.

Reopening tasks as a data steward

- You are part of the campaign and have been granted the role to access the tasks at the right state in the workflow.

  For further information, see the online publication about defining roles and workflows on Talend Help Center (https://help.talend.com).

1. In the **TASKS** page, click the name of the campaign which holds the tasks.
2. On the menu bar, use the menu combinations **RESOLVED** > **UNASSIGNED** to open the list of the tasks which are transitioned to the final state in the workflow.



3. In the panel to the right, click **ROW** and expand **Reopen tasks**.
4. From the **Selection** list, select to reopen all the tasks or only to reopen the tasks you select in the list. If you have set a filter on the tasks, you can select to reopen the filtered tasks.
5. From the **Target state** list, select:

   - **New** to send the tasks back to the initial state, or

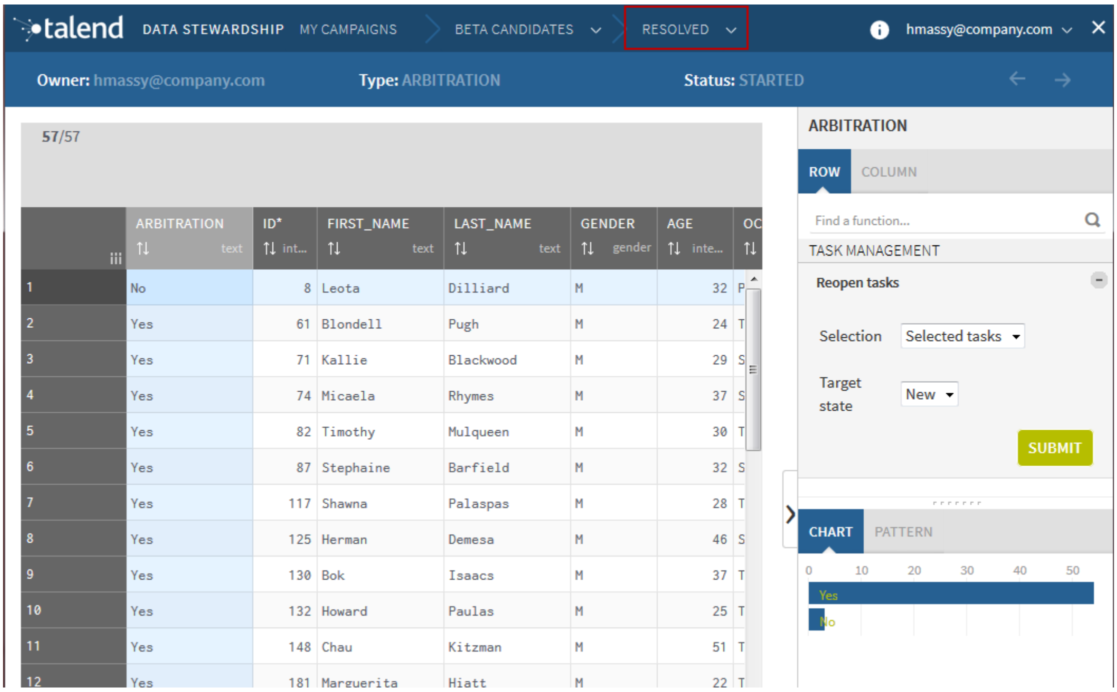
- **To validate** to send the tasks back to the second state in the workflow.

  The **To validate** option is available only if the workflow has three states.
6. Click **SUBMIT**.
   The tasks are removed from the list and transitioned back to the initial state in the workflow or to the validation state.

Reopening tasks as a campaign owner

- You have been assigned a campaign owner role in Talend Administration Center.
- You have accessed Talend Data Stewardship as a campaign owner.

1. In the **CAMPAIGNS** page, click the campaign which holds the tasks you want to reopen.
2. On the menu bar, click **NEW** and select **Resolved** to open the list of the tasks which are transitioned to the final state in the workflow.



3. In the panel to the right, click **ROW** and expand **Reopen tasks**.
4. From the **Selection** list, select to reopen all the tasks or only to reopen the tasks you select in the list. If you have set a filter on the tasks, you can select to reopen the filtered tasks.
5. From the **Target state** list, select:
   - **New** to send the tasks back to the initial state, or
   - **To validate** to send the tasks back to the second state in the workflow.

   The **To validate** option is available only if the workflow has three states.
6. Click **SUBMIT**.

The tasks are removed from the list and transitioned back to the initial state in the workflow or to the validation state based on some rules:

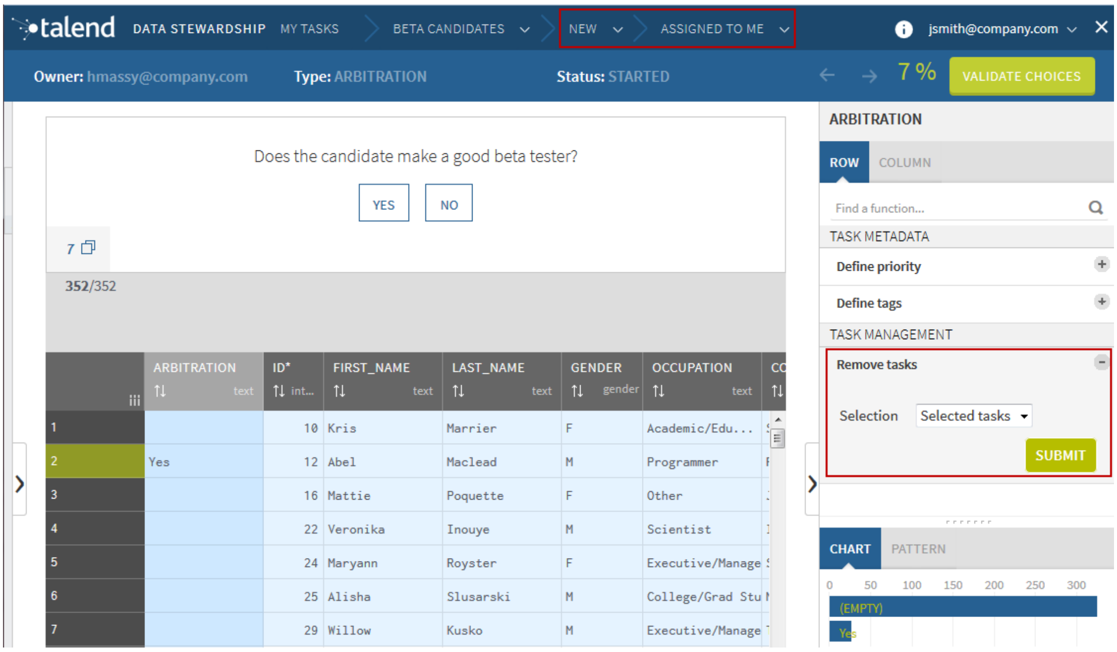
- if the current participant is allowed to access this state, then the tasks remain assigned to him,
- if the target state has only one participant, then the tasks remain assigned to this unique participant,
- otherwise, the tasks are unassigned.

Removing tasks

Data stewards can remove the tasks assigned to them in a campaign before they do any modifications on them, or after they transitioned them to the final state on the workflow.

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

1. In the **TASKS** page, click the name of the campaign which holds the tasks you want to delete.
2. On the menu bar, use the menu combinations **NEW** > **ASSIGNED TO ME** or **RESOLVED** > **ASSIGNED TO ME** to open the new tasks assigned to you or the tasks which you have transitioned to the final state respectively.



3. In the panel to the right, click **Row** and expand **Remove tasks**.
4. From the **Selection** list, select to remove all the tasks or only the tasks you select in the list.
5. Click **SUBMIT**.

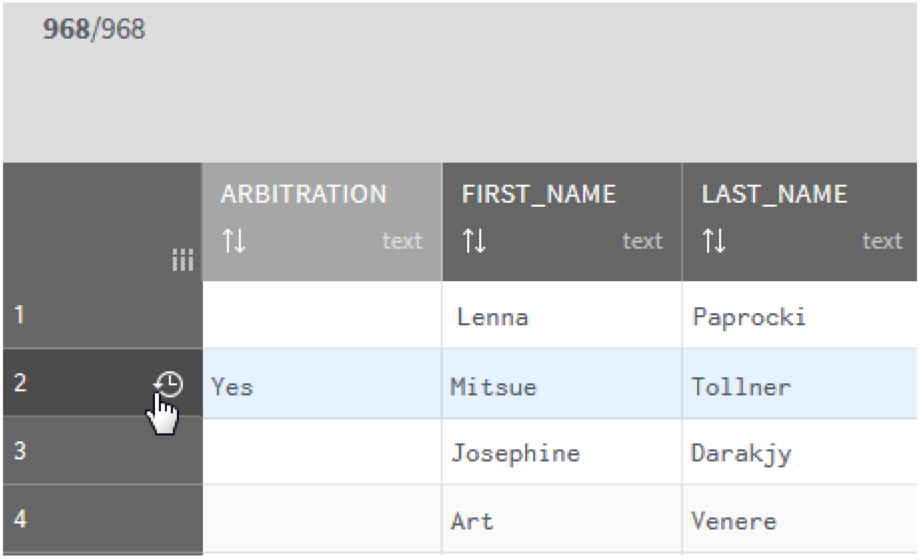
The tasks are removed from the list.

Accessing task history

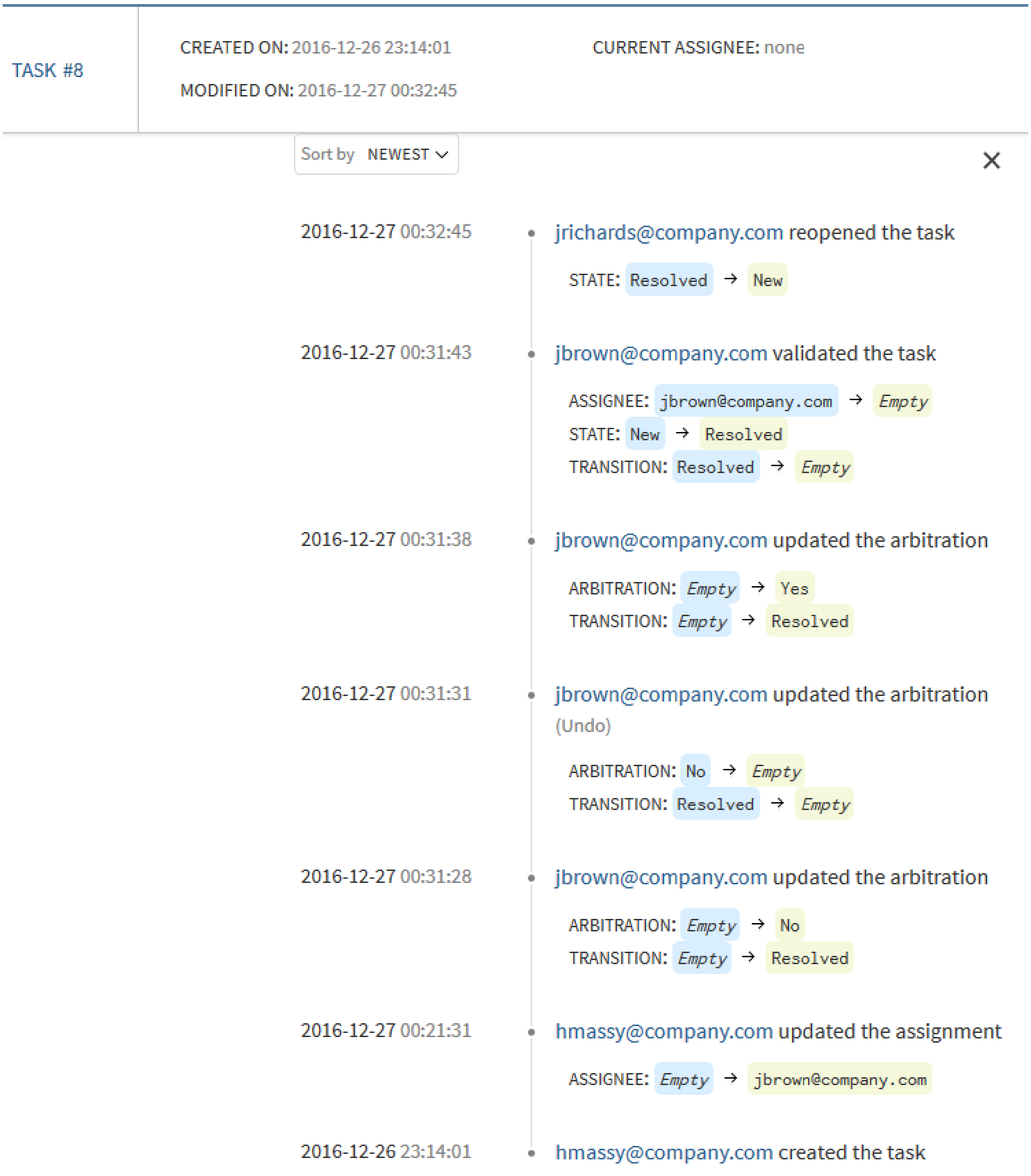
You can access the history of each of the tasks in the task list and have details about all the operations done on it and by whom.

- You have opened the task list of a campaign which you are part of.

1. In the task list, put your pointer in the field which corresponds to the task in the first column to display an icon.



2. Click the task history icon to switch to the task details.



3. Use the bar on top of the list to sort task history by dates.

4. Click the X sign to the right to go back to the task list.

List of functions on tasks

This table defines all the functions you can perform on tasks in Talend Data Stewardship.

Name	Category	Description
Arbitrate tasks Available only in **Arbitration** campaigns	task metadata	Set an arbitration choice to a group of tasks or to all tasks
Assign tasks	task metadata	Assign a group of tasks or all tasks to a specific data steward

Name	Category	Description
Define due date Available only for campaign owners	task metadata	Define the resolution delay for a group of tasks or for all tasks
Define priority	task metadata	Indicates the level of importance to a group of tasks or to all tasks
Define tags	task metadata	Adds information on a group of tasks or on all tasks to help transitioning or resolving them
Mark tasks as ready Available only in **Resolution** or **Merging** campaigns	task management	Marks a group of tasks or all tasks as ready to be validated
Remove tasks	task management	Deletes a group of tasks or all tasks from the task list
Merge the selected tasks Available only in **Merging** or **Grouping** campaigns	task management	Merges the source records of two or more tasks in a Merging campaign into a single task
Split the task Available only in **Merging** or **Grouping**campaigns	task management	Splits the source records of a given task in a Merging campaign into a separate task

Managing due dates for tasks

Campaign owners can define due dates when creating the campaigns in Talend Data Stewardship, or when loading tasks in the campaign using a Talend Job. This helps them track the most urgent tasks in the campaign and ensure that they respect service-level agreement (SLAs) they could have for task resolution.

It is also possible to define due dates after loading the tasks in the campaign. For further information, see Defining due dates for single and multiple tasks on page 64.

Due dates are always rendered in the local time zone of users. When tasks are not resolved before the due dates, the due date values are displayed in red in the **DUE DATE** column in the task list.

Defining due dates in the campaign

Campaign owners can define task due dates at campaign creation. All new tasks will be automatically labeled with the due date and time using the format `YYYY-MM-DD HH:mm:ss`. It is also possible to redefine due dates when editing the campaign.

You can then use the due dates to filter the tasks in a campaign and facilitate task priority management.

Defining due dates at campaign creation

1. Log in to Talend Data Stewardship as a campaign owner.
2. Click **CAMPAIGNS** > **ADD CAMPAIGN**.
3. Start defining the campaign parameters.

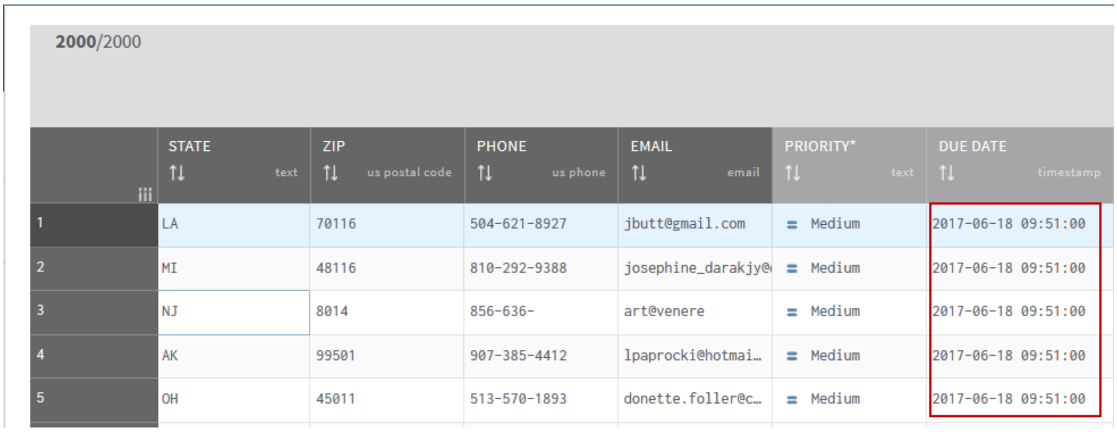
For further information, see the online publication about adding campaigns on Talend Help Center (https://help.talend.com).

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.

   The due date in a **Merging** campaign is only on master records.

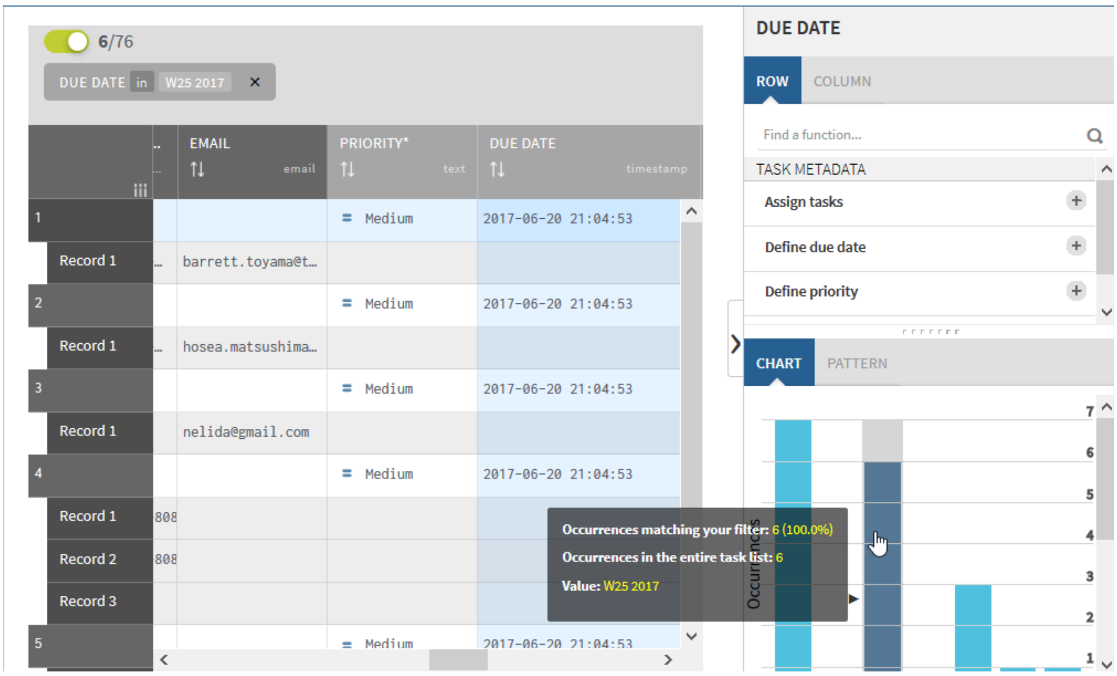
5. Finalize defining the campaign metadata and click **ADD CAMPAIGN**.
6. Use a Talend Job to load tasks in the campaign.



The due date is automatically calculated according to the delay you set in the **Delay** field.

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

7. In the task list, select the **Due Date** column and a bar in the chart to restricts the number of the listed tasks and prioritize the work on them.



For more information about different ways of filtering tasks, see Filtering tasks on page 71.

Redefining due dates when editing the campaign

Campaign owners can redefine task due dates when editing the campaign. All tasks created in the campaign after this change will be automatically labeled with the new due date and time using the format `YYYY-MM-DD HH:mm:ss`. However, old tasks listed in the campaign will keep the old delay.

1. Log in to Talend Data Stewardship as a campaign owner.
2. Point to the campaign in which you want to modify the due date and click the gear icon which is displayed to the right of the campaign name.
3. Change the value in the **Delay** field.
4. Finalize editing the campaign metadata and click **EDIT CAMPAIGN**.
5. Use a Talend Job to load new tasks in the campaign.

   The due date is automatically calculated according to the new delay you set in the **Delay** field. Initial tasks keep the old due date, while all new tasks are labeled with new due date values.

   For further information, see the online publication on writing tasks in campaigns on Talend Help Center (https://help.talend.com).
6. In the task list, select the **Due Date** column and a bar in the chart to restricts the number of the listed tasks and prioritize the work on them.

   For more information about different ways of filtering tasks, see Filtering tasks on page 71.

Defining due dates from Talend studio

Campaign owners can define task due dates when loading the tasks in the campaign by using a Talend Job. Running the Job creates the tasks in the campaign and label them with the due date and time using the format `YYYY-MM-DD HH:mm:ss`.

However, setting a due date when loading tasks overrides the value which you may have defined when creating or editing the campaign.

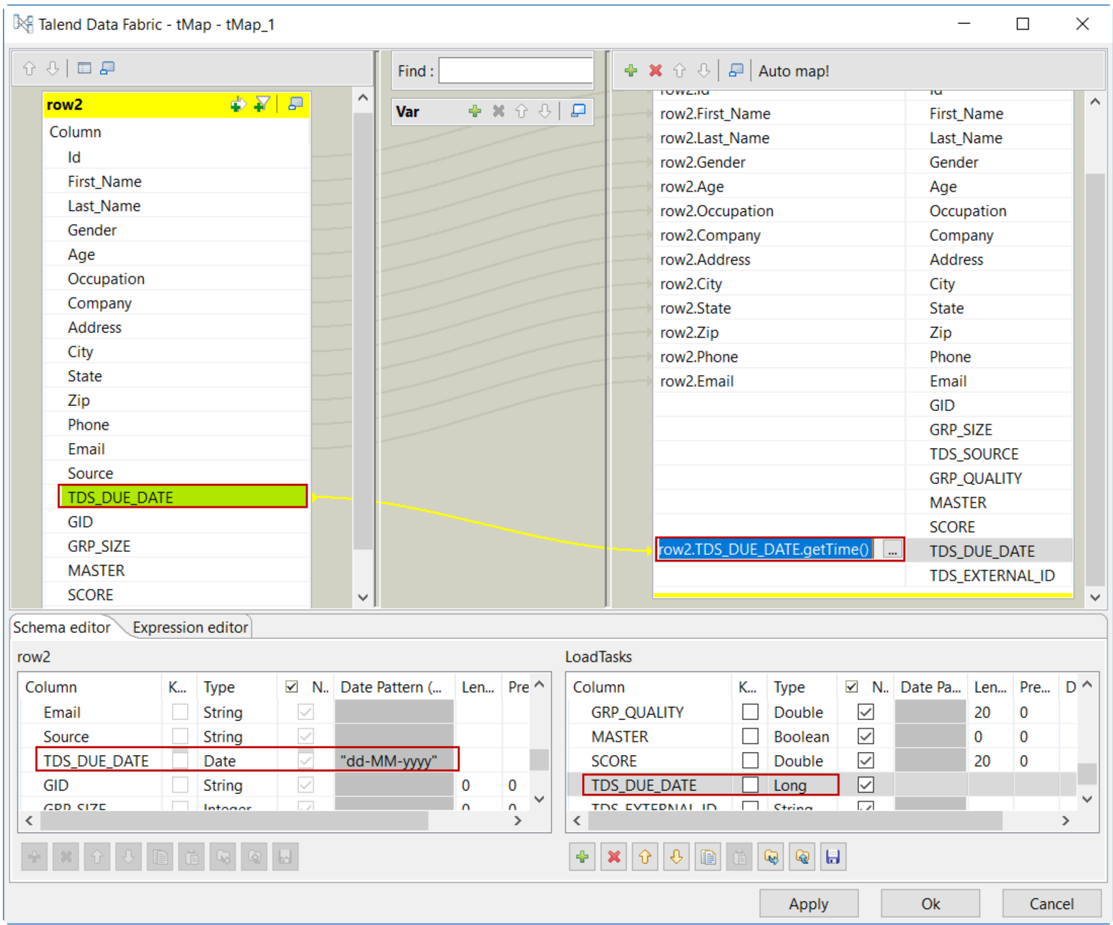
You can then use the due dates to filter the tasks in a campaign and facilitate task management.

1. Create a Job with the **tMap** and **tDataStewardshipTaskOutput** components to load data from a file into a campaign you have already created in Talend Data Stewardship.

   Suppose that the input data flow has a due date column of the type `Date` for this example.

   When computing the due date in the Job, you can consider two main use cases :

   - The service-level agreement (SLA) varies based on the data. For instance, you are dealing with customer deduplication, you have two days to solve a duplicate for Tier-1/platinum customers but five days for Tier-2/Gold customers.
   - The due date is computed via an external calendar to include days off, i.e. week-ends, public holiday, etc..
2. Use the **tMap** component to map the input data flow to the output flow and to transform the `Date` type of the due date column to `Long`.

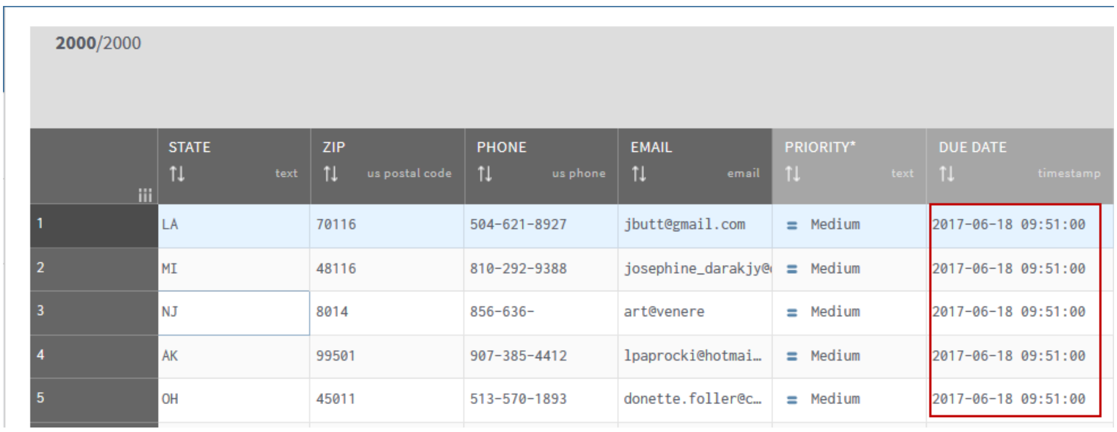


For further information, see the online publication about the **tMap** component on Talend Help Center (https://help.talend.com).

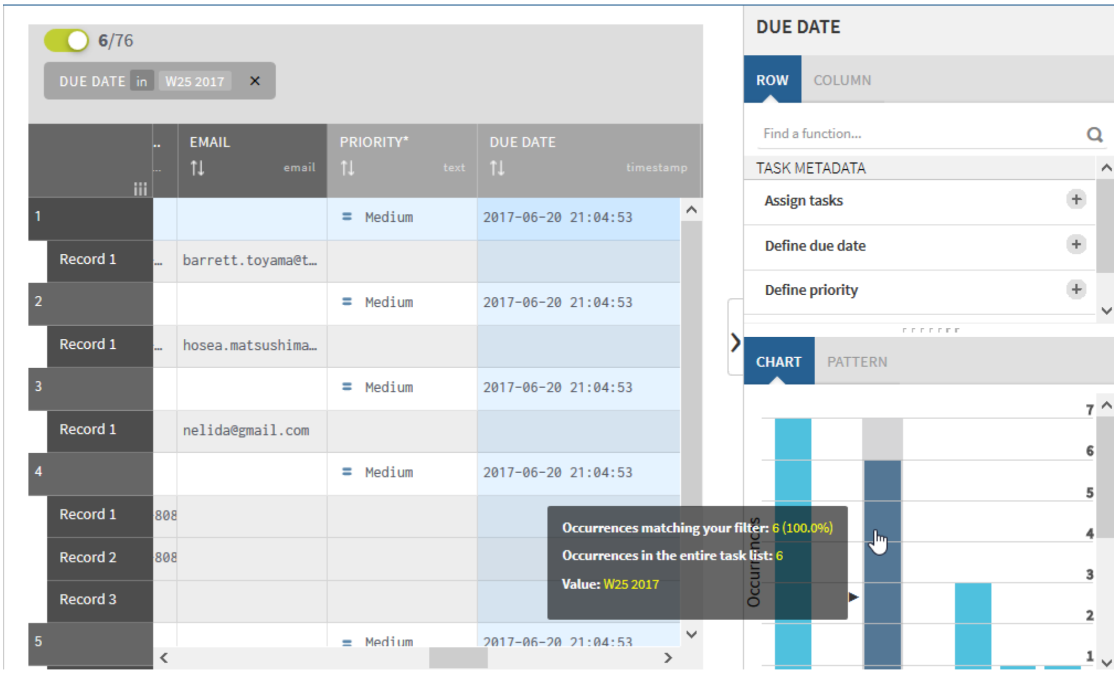
3. Configure the **tDataStewardshipTaskOnput** component to load the tasks into the campaign defined on Talend Data Stewardship.

   You can find information about the **tDataStewardshipTaskOnput** component on Talend Help Center (https://help.talend.com).

4. Run your Job to load the tasks in the campaign with the due date.



5. Select the **Due Date** column and a bar in the chart to restricts the number of the listed tasks and prioritize the work on them.



For more information about different ways of filtering tasks, see Filtering tasks on page 71.

Managing task metadata and using bulk operations

When campaign owners use Talend Jobs to load tasks in Talend Data Stewardship, they can set priority levels, define due dates or add tags for tasks. Otherwise, they can do the same from within the task list.

Indexing tasks with priority level, due dates and/or tags can help bringing more structure to the process of handling tasks. It also enables you to filter the tasks accordingly using the chart in the right panel. You can then for example bulk assign, bulk delegate or bulk resolve the listed tasks.

For a list of all available functions on tasks, see List of functions on tasks on page 59.

Assigning multiple tasks

In Talend Data Stewardship, campaign owners and/or data stewards who are part of a specific campaign can bulk assign the tasks listed in the campaign.

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

1. In the home page, select **CAMPAIGNS** if you are a campaign owner and **TASKS** if you are a data steward
2. Click the name of the campaign which holds the tasks you want to assign.
3. In the panel to the right, click **ROW** and expand **Assign tasks**.
4. From the **Selection** list, select to assign all the tasks or only to the tasks you select in the task list.
5. From the **Assignee** list, select the data steward to whom to assign the tasks and click **SUBMIT**. This list shows all the data stewards defined in the campaign.

The selected tasks are assigned to the data steward.

Defining due dates for single and multiple tasks

From the task list, campaign owners can define due dates to single tasks manually or perform the update on multiple tasks from the task list.

Due dates help you track the most urgent tasks in the campaign.

- A campaign owner has created the campaign and loaded tasks into it using a Talend Job.

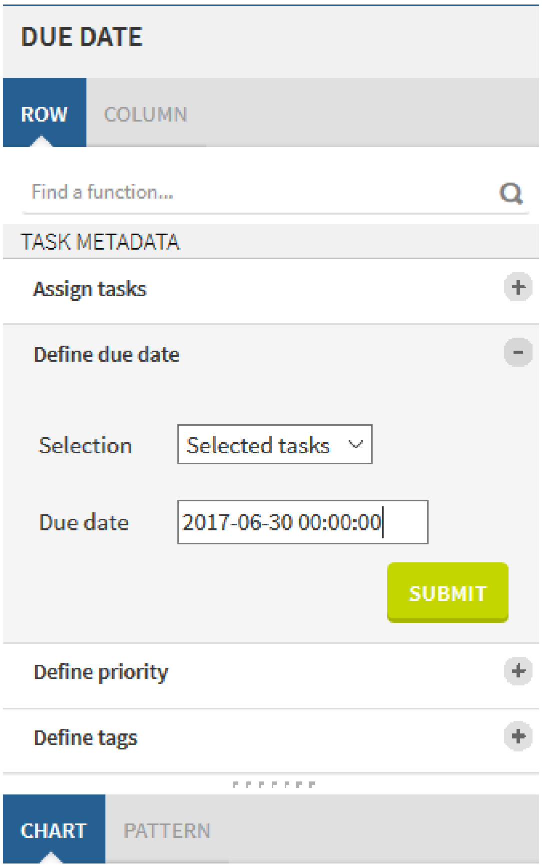
  For further information, see the online publication about writing tasks in campaigns on Talend Help Center (https://help.talend.com).

Define due dates for single tasks

1. Log in to Talend Data Stewardship as a campaign owner and click the campaign which holds the tasks you want to update.
2. Browse to the task and double-click the field in the **DUE DATE** column. In merging or grouping tasks, double-click the master record.
3. Enter the date and time for the task using the format: `YYYY-MM-DD HH:mm:ss` and press **Enter**. The due date should follow the exact timestamps format.
4. Repeat the above steps to define or modify due dates for more tasks.

Define a due date for multiple tasks

1. Log in to Talend Data Stewardship as a campaign owner and click the campaign which holds the tasks you want to update.
2. In the panel to the right, click **ROW** and expand **Define due date**.



3. From the **Selection** list, select to add the due date timestamps to all the tasks or only to the tasks you select in the task list.
4. In the **Due date** field, enter the date and time using the format: `YYYY-MM-DD HH:mm:ss`. The due date should follow the exact timestamps format.
5. Click **SUBMIT** to update the tasks in the open list with the due date.

Defining priorities for single and multiple tasks

When tasks are created in a campaign, priority level is automatically set by default to **Medium** if the campaign owner does not set a priority value in the Job properties. Priority levels indicate the level of importance you want to give to the tasks.

Several priority levels are available including **Very high**, **High**, **Medium**, **Low** and **Very low**. You can manually define or change the priority level of one or multiple tasks from within Talend Data Stewardship according to your needs.

- A campaign owner has created the campaign and granted you access to it.
- A campaign owner has loaded the tasks in the campaign using a Talend Job.
- Tasks have been assigned to data stewards.

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

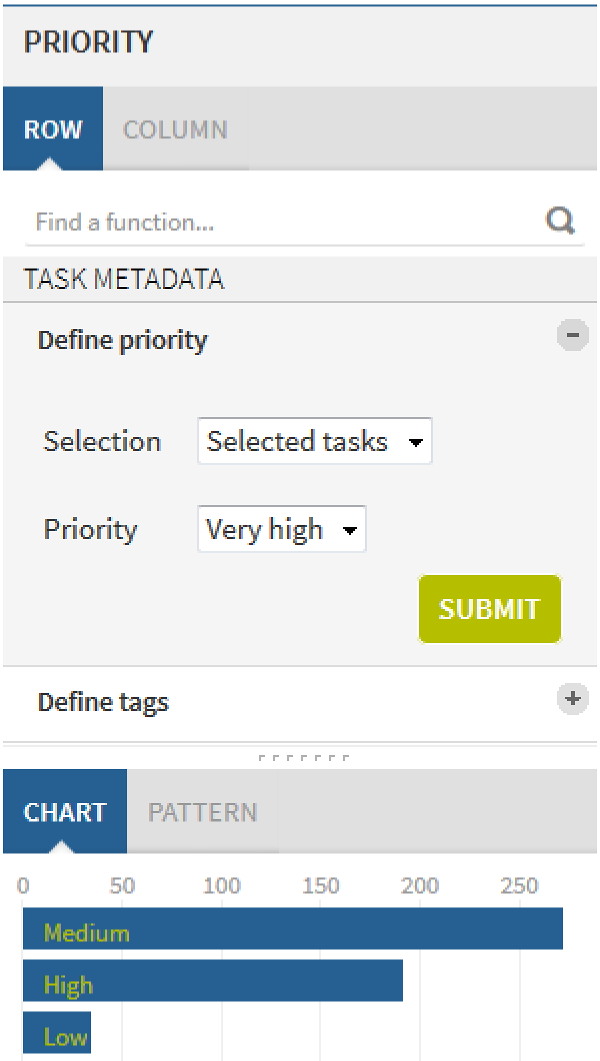
Define a new priority value for single tasks

1. Depending on the role assigned to you in Talend Data Stewardship:
   - Log in as a campaign owner and in the **Campaign** page, point to the campaign which holds the tasks you want to manage and click ,
   - Log in as a data steward and in the **Tasks** page, click the campaign which holds the tasks you want to manage.
2. Browse to the task and double-click the field in the **PRIORITY** column.
3. From the drop down list, select the priority level you want to give to the task.
4. Repeat the above steps to define priority levels to more tasks.

The new priority level of the task is defined in the **PRIORITY** column.

Define a new priority value for multiple tasks

1. Depending on the role assigned to you in Talend Data Stewardship:
   - Log in as a campaign owner and in the **Campaign** page, point to the campaign which holds the tasks you want to manage and click ,
   - Log in as a data steward and in the **Tasks** page, click the campaign which holds the tasks you want to manage.
2. In the panel to the right, click **ROW** and expand **Define priority**.



3. From the **Selection** list, select to add the new priority value to all the tasks or only to the tasks you select in the task list.
4. From the **Priority** list, select the new priority value and click **SUBMIT**.

The new priority level is defined in the **PRIORITY** column according to the parameters you set.

Defining tags for single and multiple tasks

When tasks are created in a campaign, a campaign owner can automatically add one or multiple tags to these tasks. Tags are intended to be words or short phrases which help adding some information on tasks to help transitioning or resolving them.

However, you can modify tags or add new ones manually to one or multiple tasks from within Talend Data Stewardship according to your needs.

- A campaign owner has created the campaign and granted you access to it.
- A campaign owner has loaded the tasks in the campaign using a Talend Job.
- Tasks have been assigned to data stewards.

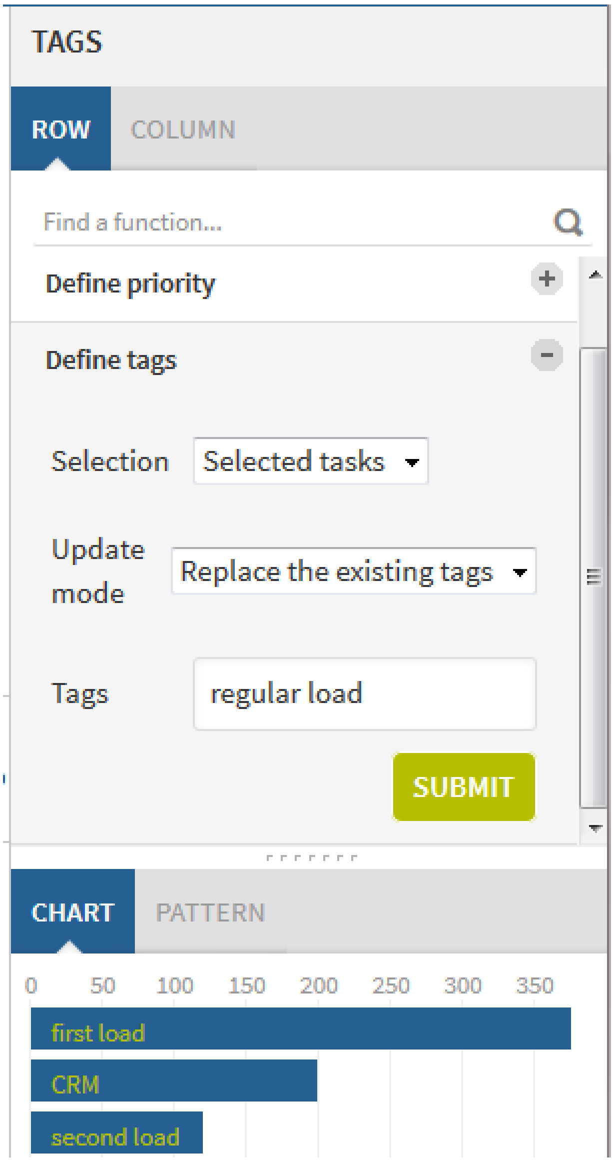
  For further information, see the online publication about assigning tasks on Talend Help Center (https://help.talend.com).

Define or modify tags for single tasks

1. Depending on the role assigned to you in Talend Data Stewardship:
   - Log in as a campaign owner and in the **Campaign** page, point to the campaign which holds the tasks you want to manage and click ,
   - Log in as a data steward and in the **Tasks** page, click the campaign which holds the tasks you want to manage.
2. Browse to the task and double-click the field in the **TAGS** column.
3. Modify the tag or the tags which have been already added to the task, or add completely new tags. Use commas to separate multiple tags.
4. Repeat the above steps to modify or add new tags to more tasks in the list.

Define or modify tags for multiple tasks

1. Depending on the role assigned to you in Talend Data Stewardship:
   - Log in as a campaign owner and in the **Campaign** page, point to the campaign which holds the tasks you want to manage and click ,
   - Log in as a data steward and in the **Tasks** page, click the campaign which holds the tasks you want to manage.
2. In the panel to the right, click **ROW** and expand **Define tags**.



3. From the **Selection** list, select to add the new tag(s) to all the tasks or only to the tasks you select in the task list.
4. From the **Update mode** list, select to add the new tag(s) already defined or to completely replace them and click **SUBMIT**.

In the task list, the tags in the **TAGS** column are modified according to the parameters you set.

Selecting arbitration choices for multiple tasks

When handling tasks in an **Arbitration** or a **Grouping** campaign, you can as a data steward select one arbitration choice for multiple tasks at once and transition the tasks to the second state defined in the workflow.

- A campaign owner has created the campaign and granted you access to it.
- A campaign owner has loaded the tasks in the campaign using a Talend Job.
- Tasks have been assigned to data stewards.

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

1. Log in as a data steward and in the **Tasks** page, click the **Arbitration** or the **Grouping** campaign which holds the tasks you want to manage.
2. In the panel to the right, click **ROW** and expand **Arbitrate tasks**.
3. From the **Selection** list, select to set the choice to all the tasks or only to the tasks you select in the task list.
   Available choices are those defined while creating the campaign in addition to one extra default option: **Clear arbitration choice**.
4. From the **Arbitration** list, select what choice to set to the tasks and click **SUBMIT**.
5. To reset any arbitration choice made so far on the selected tasks, select **Clear arbitration choice** from the list and click **SUBMIT** .

The selected tasks have now the arbitration choice you set and the task list is reloaded.

List of functions on tasks

This table defines all the functions you can perform on tasks in Talend Data Stewardship.

Name	Category	Description
Arbitrate tasks Available only in **Arbitration** campaigns	task metadata	Set an arbitration choice to a group of tasks or to all tasks
Assign tasks	task metadata	Assign a group of tasks or all tasks to a specific data steward
Define due date Available only for campaign owners	task metadata	Define the resolution delay for a group of tasks or for all tasks
Define priority	task metadata	Indicates the level of importance to a group of tasks or to all tasks
Define tags	task metadata	Adds information on a group of tasks or on all tasks to help transitioning or resolving them
Mark tasks as ready Available only in **Resolution** or **Merging** campaigns	task management	Marks a group of tasks or all tasks as ready to be validated
Remove tasks	task management	Deletes a group of tasks or all tasks from the task list
Merge the selected tasks Available only in **Merging** or **Grouping** campaigns	task management	Merges the source records of two or more tasks in a Merging campaign into a single task
Split the task Available only in **Merging** or **Grouping**campaigns	task management	Splits the source records of a given task in a Merging campaign into a separate task

Filtering tasks

Filtering tasks in Talend Data Stewardship restricts the number of the listed tasks and facilitates any further processing. You can filter tasks using the quality bar displayed on top of the task list, the charts displayed next to the task list, or the **PATTERN** tab in the right panel.

You can filter the tasks in a campaign for which you are granted the access as a campaign owner or the tasks assigned to you as a data steward.

Filtering tasks using the quality bar

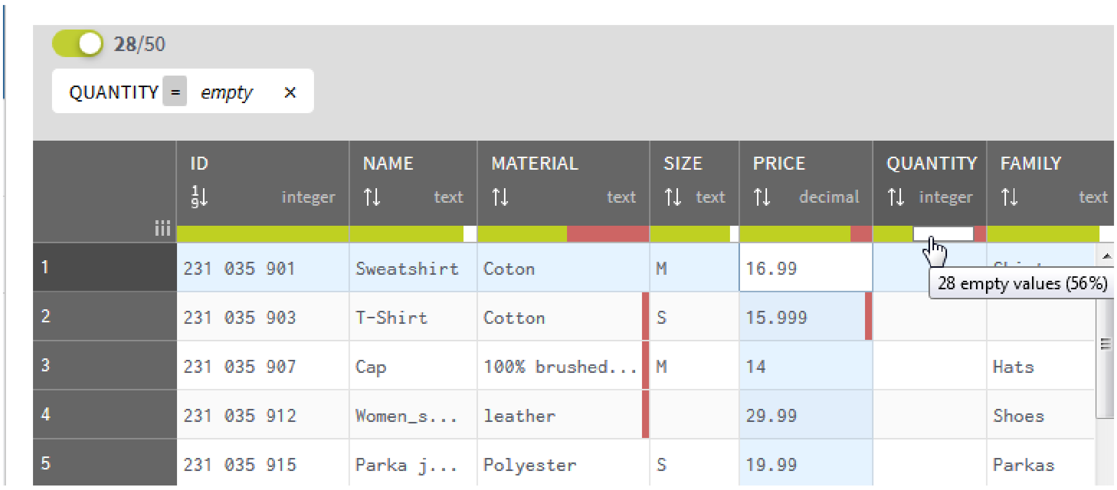
You can use the quality bar to filter data and list only the tasks you want to resolve or you want to assign or delegate to specific data stewards.

The quality bar is a bar on top of each of the columns in a task list. It indicates the quality of data using colors. The quality bar gives information about how many valid, invalid or empty values are there in the column. The quality bar is available in all campaign types except an **Arbitration** campaign.

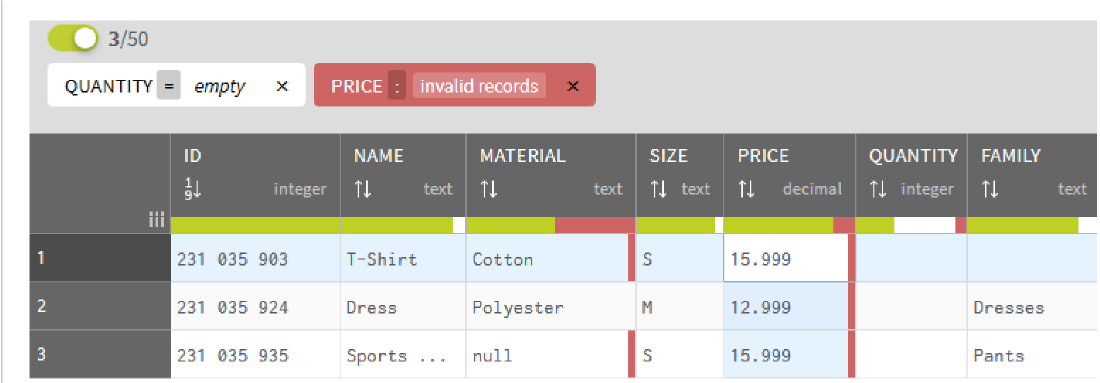
- You have opened tasks in a **Resolution** or **Merging** campaign.

1. Point to a color in the quality bar on top of one of the columns to display the number of the values which match the color indication.

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



2. Click the color in the quality bar of a column to list only valid, invalid or empty fields in that column.
   The filter detail is added on top of the list using the same filter color, and a switch is displayed on the top left corner.
3. Click a color in the quality bar of another column to combine the two filters and list the tasks that satisfy the combined conditions.
   You can combine as many filters as needed and you can combine filters between the quality bar and the chart in the right-hand panel.



4. Click the switch to toggle between the filtered and unfiltered task list, if needed.
5. To remove the filter(s) you defined, place your pointer on the top right corner of the list and click the trash icon.

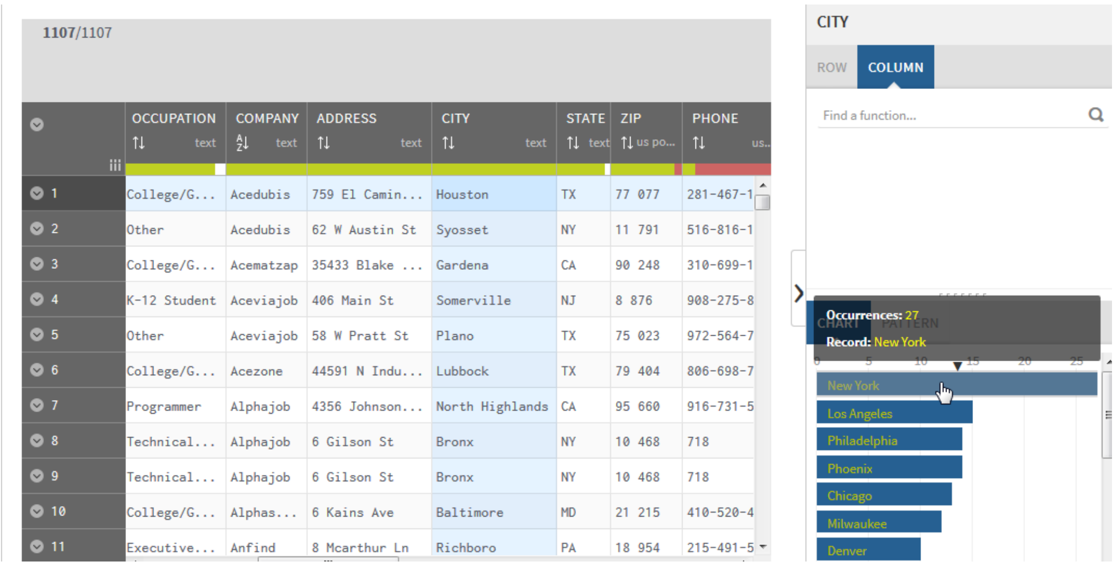
Filtering tasks using charts

You can use the charts displayed next to the task list to filter data and list only the task you want to resolve or you want to assign or delegate to specific data stewards.

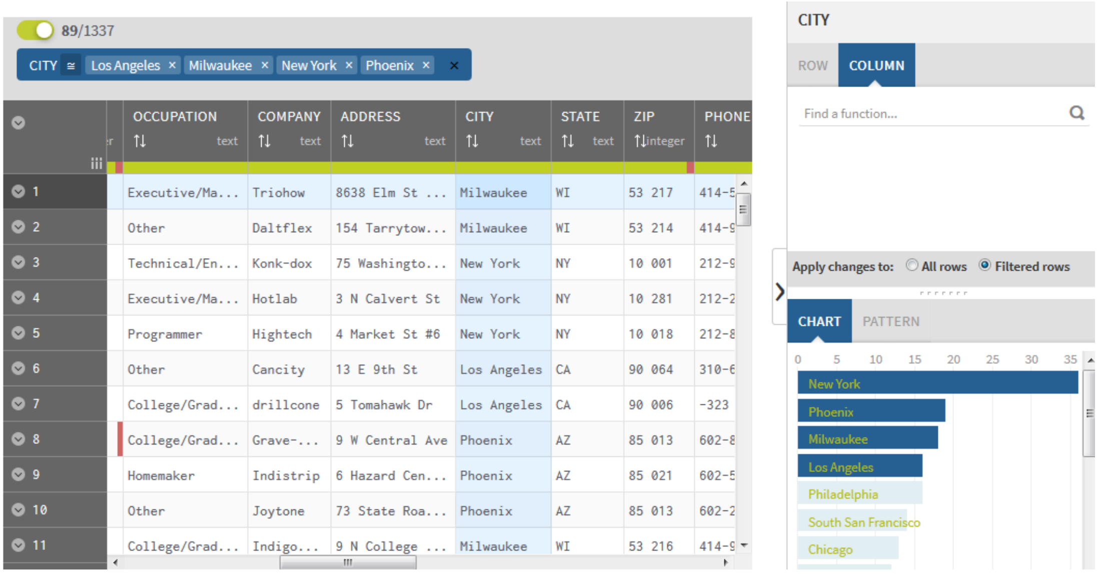
Charts are graphical representations of the distribution of the values contained in the selected column. Charts are displayed next to the task list and they differ according to the type of the data you select.

- You have opened a task list in one of the campaign defined in Talend Data Stewardship.

1. Click the header of a column in the task list to select its content.
   The distribution of the values in the columns is represented by a vertical bar chart for text values and by horizontal bar chart for numerical values.



2. Point to any bar in the chart to display the value name the number of its occurrences.
3. Click the bar which values you want to filter, or hold the SHIFT key and select multiple bars in the chart to list the corresponding tasks.
   The filter detail is added on top of the list and a switch is displayed on the top left corner.



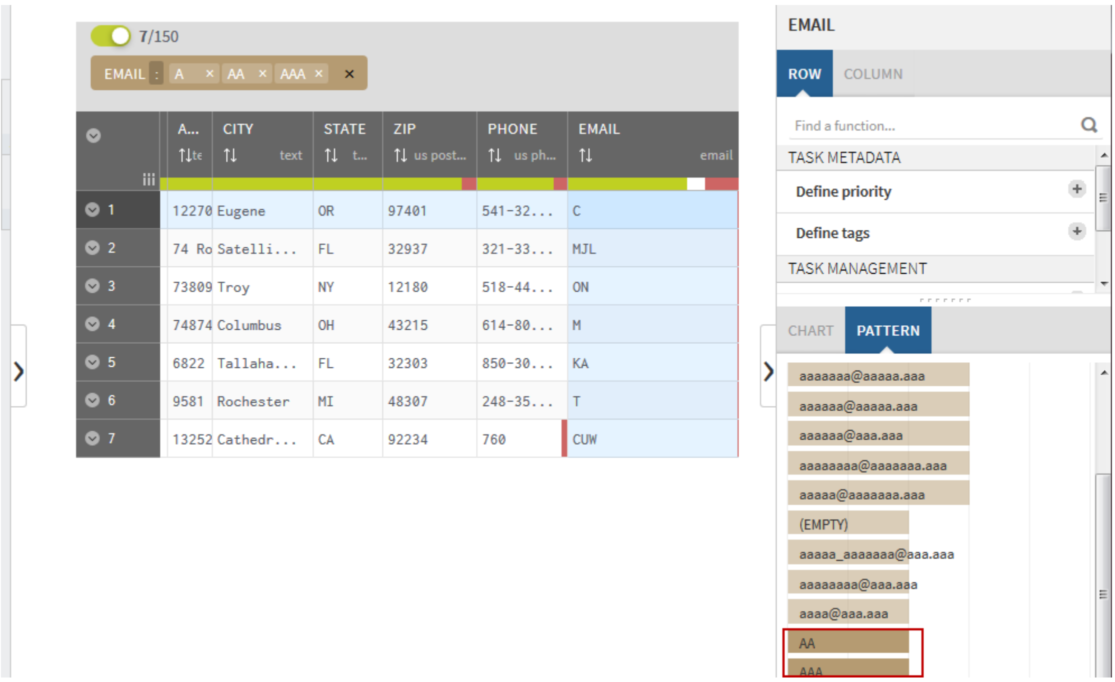
4. Click a color in the quality bar of a column to combine filters from the charts with filters from the quality bar and list the tasks that satisfy the combined conditions, if needed.
5. Click the switch to toggle between the filtered and unfiltered task list, if needed.
6. To remove the filter(s) you defined, place your pointer on the top right corner of the list and click the trash icon.

Filtering tasks using patterns

You can use the **PATTERN** tab in the right-hand panel to filter data and highlight and list only incorrect, unusual, or unexpected values.

- You have opened a task list in one of the campaign defined in Talend Data Stewardship.

1. Click the header of a column in the task list to select its content, the `EMAIL` column in this example.
2. Select **PATTERN** in the right-hand panel.
   All patterns which represent the values in the selected column are displayed in the right panel.



3. Click the pattern which values you want to filter, or hold the SHIFT/Ctrl key and select multiple patterns to list the corresponding tasks.
   The filter detail is added on top of the list and a switch is displayed on the top left corner.
4. Click the switch to toggle between the filtered and unfiltered task list, if needed.
5. To remove the filter(s) you defined, place your pointer on the top right corner of the list and click the trash icon.

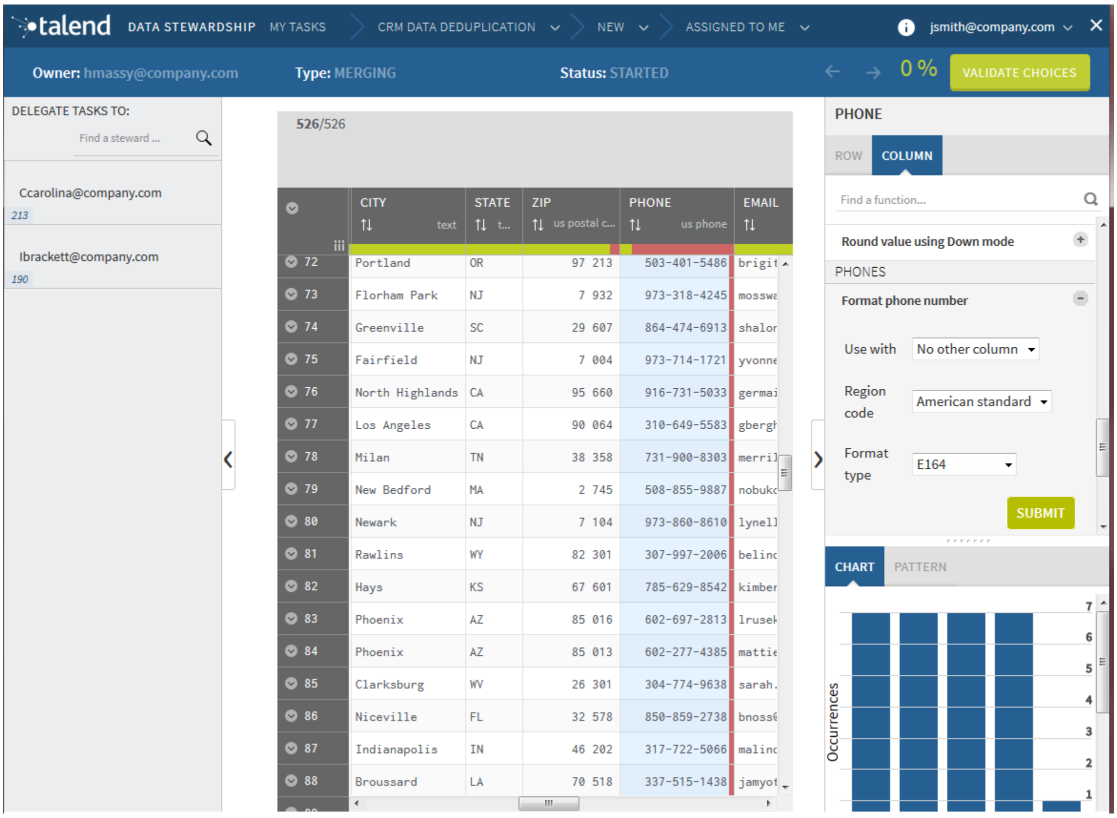
Transforming data in a column

When handling tasks in a **Resolution** or **Merging** campaign, data stewards can use a list of different functions on a given column in the task list and do a bulk transformation of the data in the selected column.

- A campaign owner has defined you as part of a **Resolution** or **Merging** campaign.
- A campaign owner has assigned you tasks in the campaign. Otherwise, you can assign the tasks to yourself.

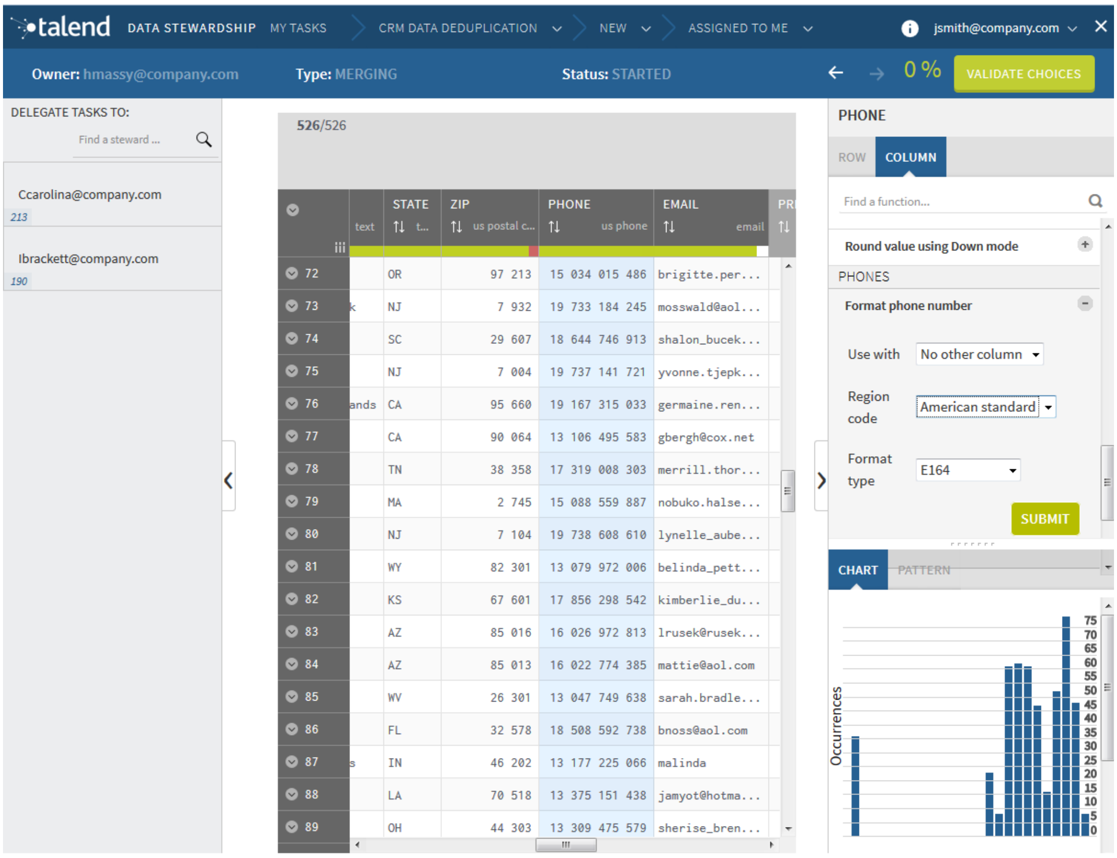
  For further information, see the online publication about assigning tasks on Talend Help Center (https://help.talend.com).

1. In the **TASKS** page, click the name of the campaign you want to open.
2. In the task list, click the heading of the column for which you want to do the bulk modification.
   The column name is displayed at the top of the right panel and all available functions are listed under **COLUMN**.



3. Browse the list to reach the function you want to use on the column data, or type the function name in the search field.
   In this example, phone numbers are not conformed to the general format of the international phone numbers. For further information about the functions in the right panel, see List of functions on columns on page 75.

4. Expand **Format phone number** and:
   a) From the **Region code** list, select **American standard**.
   b) From the **Format type** list, select **E164**.
5. Click **SUBMIT**.



When you use functions on columns in a **Merging** campaign, only values in the master fields are transformed, i.e. values in the source fields are not

Phone numbers in the column are transformed to conform to the general standard of the American phone numbers.

List of functions on columns

This table defines all the functions you can perform on columns in Talend Data Stewardship.

Name	Category	Description
Negate value	boolean	Reverses the boolean value of cells from this column
Clear on matching value	data cleansing	Clears cells that match the value
Clear the cells with invalid values	data cleansing	Clears cells that contain a value recognized as invalid
Delete the rows with empty cell	data cleansing	Deletes rows that have empty cells

Name	Category	Description
Fill cells with value	data cleansing	Fills cells from this column with a given value
Fill empty cells with text	data cleansing	Fills empty cells from this column with a given value
Fill invalid cells with value	data cleansing	Fills cells from this column with a given value
Convert date	dates	Convert date measurement units
Modify date	dates	Adds or substracts time unit amount
Convert distance	conversions	Convert distance measurement units
Convert duration	conversions	Convert time duration measurement units
Calculate absolute value	math	Calculates the absolute value for all the numeric values in this column.
Remove fractional part	numbers	Round towards zero. (3.74 -> 3) and (-3.74 -> -3)
Round value using ceil mode	numbers	Rounds value to the largest integer (3.14 -> 4 if **Precision** is set to `0`, and 3.14 -> 3.2 if **Precision** is set to `1`)
Round value using floor mode	numbers	Rounds value to the smallest integer (3.74 -> 3 if **Precision** is set to `0`, and 3.74 -> 3.7 if **Precision** is set to `1`)
Round value using halfup mode	numbers	Rounds value to the closest integer (3.14 -> 3 and 3.74 -> 4 for a **Precision** set to `0`)
Round value using down mode	numbers	Rounds towards zero. (3.74 -> 3 and -3.74 -> -3 for a **Precision** set to `0`)
Format phone number	phones	Formats a phone number to the selected format
Change to lower case	strings	Converts all of the cell text in this column to lower case

Name	Category	Description
Change to title case	strings	Converts the text content from this column to title case (i.e. "data prep" -> "Data Prep")
Change to upper case	strings	Converts all of the cell text in this column to UPPER case (capitalize)
Remove part of the text	strings	Removes specified text from cells in this column
Remove consecutive characters	strings	Removes repeated characters (i.e. spaces)
Remove trailing and leading characters	strings	Removes trailing and leading spaces (i.e. trim)
Replace the cells that match	strings	Replaces the cells that have a specific value
Add extra characters	strings advanced	Adds extra characters (padding) on the left or on the right of the original value to match an expected size
Remove all non alpha numeric characters	strings advanced	For example ÃƒÂ¢Ã¢Â €ÂšÃ‚Â¬10.5k will become 105
Remove all non numeric characters	strings advanced	For example ÃƒÂ¢Ã¢Â €ÂšÃ‚Â¬10.5k will become 10.5
Simplify text (remove case, accent, etc.)	strings advanced	Simplifies the content of this column (i.e. François -> francois)
Apply survivorship rule	survivorship	Decides what value to use to build the master record

Managing campaigns

Managing campaigns

In Talend Data Stewardship, campaign owners can perform some basic management procedures on campaigns.

They can edit or delete the campaigns for which they have access rights.

Editing a campaign

You can edit the campaigns that have been created in Talend Data Stewardship whether they are started or not started yet. However the options available when editing a campaign which has started are limited.

- The campaign has been defined in Talend Data Stewardship and you are defined as the campaign owner.
- You have accessed Talend Data Stewardship as a campaign owner.

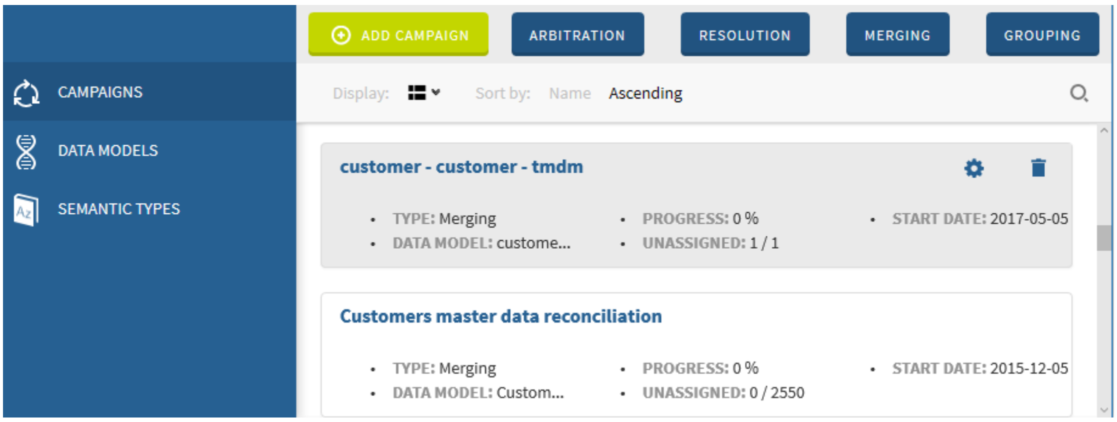
1. In the home page, click **CAMPAIGNS** to open the list of the campaigns for which you are the owner.
2. Point to the campaign you want to edit and click the gear icon displayed to the right of the campaign name.
   You can not modify the name and type of the campaign, the data model used in the campaign and the workflow states.
3. Add new campaign owners or data stewards or modify the roles as needed.
4. Modify the access rights and survivorship rules used on the attributes in the data model selected in the campaign.
5. Modify the roles assigned to the workflow states.
6. Click **EDIT CAMPAIGN** at the bottom of the page to validate your changes.

Deleting a campaign

You can delete any of the campaigns listed in Talend Data Stewardship for which you have access rights. This will delete all the tasks loaded into the campaign.

- The campaign has been defined in Talend Data Stewardship.
- You have accessed Talend Data Stewardship as a campaign owner.

1. In the home page, click **CAMPAIGNS** to open the list of the campaigns for which you are the owner.



2. Point to the campaign you want to delete and click the trash can displayed to the right of the campaign name.
3. Click **REMOVE** in the confirmation message which is displayed.

The campaign is deleted with all the tasks created in it.