



Talend Data Preparation User Guide

2.1

Contents

Copyright.....	3
Administer Talend Data Preparation.....	4
F.A.Q.....	15
Managing datasets.....	17
Managing preparations.....	21
Discovering the data.....	34
Working with the data.....	61
Developer tasks.....	106
Reference.....	118

Copyright

Adapted for 2.1. Supersedes previous releases.

Publication date: June 29th, 2017

Copyright © 2017 Talend. All rights reserved.

Notices

Talend is a trademark of Talend, Inc.

All brands, product names, company names, trademarks and service marks are the properties of their respective owners.

End User License Agreement

The software described in this documentation is provided under **Talend's** End User License Agreement (EULA) for commercial products. By using the software, you are considered to have fully understood and unconditionally accepted all the terms and conditions of the EULA.

To read the EULA now, visit <http://www.talend.com/legal-terms/us-eula>.

Administer Talend Data Preparation

What is Talend Data Preparation?

Talend Data Preparation is a self-service application that enables information workers to cut hours out of their work day by simplifying and expediting the laborious and time-consuming process of preparing data for analysis or other data-driven tasks.

The application is delivered in two different versions:

- An open source-based, personal desktop solution with 1-click install, modern user experience, and import/export capabilities for data exchanges.
- A subscription version that runs on top of the Talend Integration Platform and delivers enterprise-class capabilities together with connectivity to virtually any data source. It fosters collaboration between business people who know the data best and central organizations, like IT or Risk Management, that define the rules and policies for data accessibility and governance.

It includes:

- Integration and cataloging
- Data Discovery and Profiling
- Cleansing, standardizing and shaping
- Enriching and connecting datasets
- Operationalizing Data Preparation

Talend Data Preparation concepts

These definitions will help you understand the main concepts in Talend Data Preparation.

Dataset

A dataset holds the raw data that can be used as the raw material for one or more preparations. It is presented as a table on which you can apply recipe steps without affecting the original data. A dataset can be reused across preparations.

Function

A function is an action applied on a row or a column in a dataset such as removing empty rows. As functions are applied as part of a preparation, they do not modify the original data. Applied functions are recorded, in sequence, into recipes.

Preparation

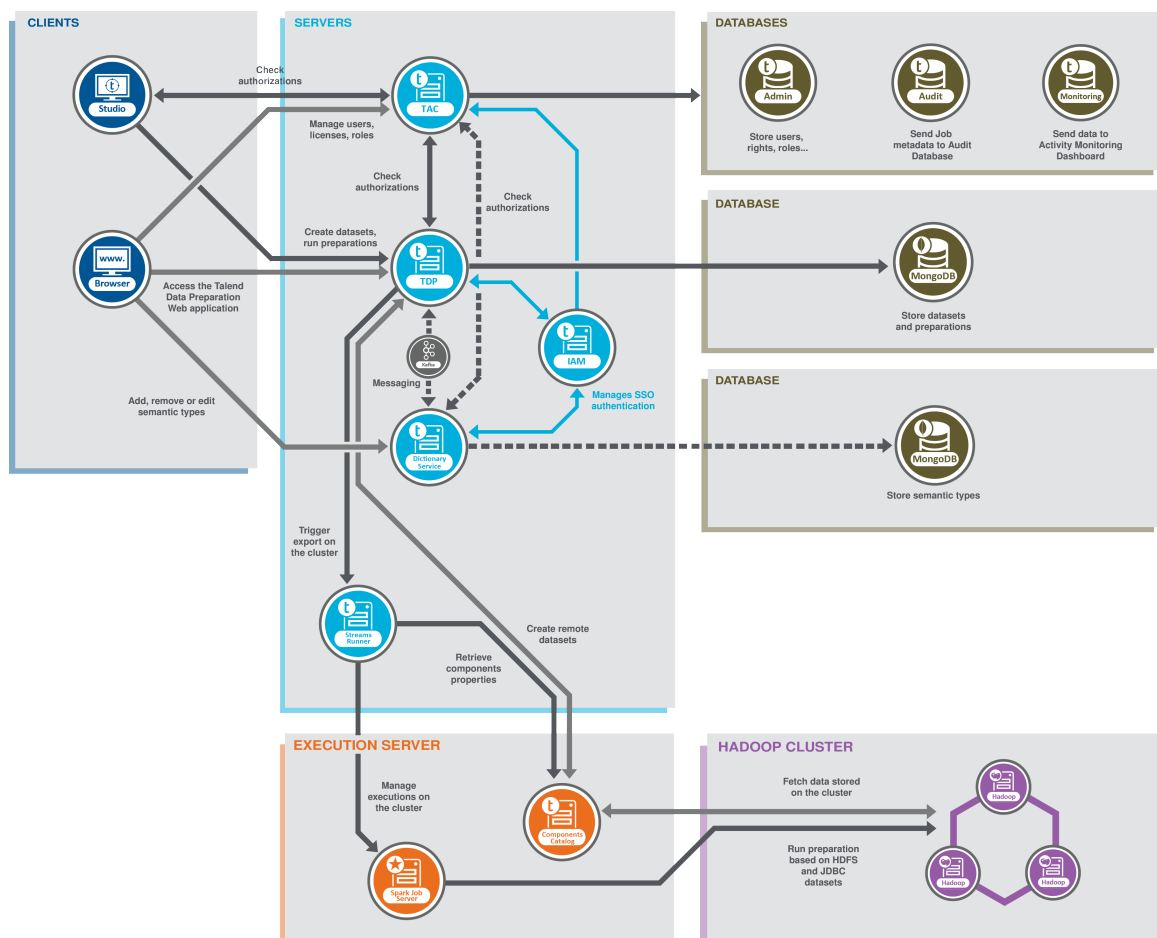
A preparation is what links a dataset and a recipe together: it is the final outcome that you want to achieve with your data. You can export this outcome as a file or connect it to data targets. A preparation takes one dataset and applies a recipe to produce an outcome. The original dataset is never modified.

Recipe

A recipe is literally defined as "a set of directions with a list of ingredients for making or preparing something". In Talend Data Preparation, the ingredients are the raw data, called datasets, and the directions are the set of functions applied to the dataset. Visually, the recipe is the top-down sequence of functions in the left collapsible panel. A recipe is linked to the dataset through a preparation. Every update of the recipe is automatically saved in the preparation all the time.

Talend Data Preparation architecture

This architecture diagram identifies the functional blocks of Talend Data Preparation, and the interactions between them.



Five different functional blocks are defined:

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

From the Web browser, you access the Talend Data Preparation Web application. This is where you import your data, from local files or other sources, and cleanse or enrich it by creating new preparations on this data. In addition, you can optionally access Talend Dictionary Service server to 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 UI](#) on page 35.

In Talend Studio, you can benefit from the Talend Data Preparation features through the use of the **tDatasetInput**, **tDatasetOutput**, and **tDataprepRun** components. You can create datasets from

various databases and export them in Talend Data Preparation, or leverage a preparation directly in a data integration Job or Spark Job.

- The **Servers** block includes the Talend Data Preparation application server, connected to Talend Administration Center, and optionally Talend Dictionary Service server and the Streams Runner server. This block also includes a Kafka server used for internal messaging between Talend Data Preparation and Talend Dictionary Service. 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 Preparation. From there, it is also possible to execute Jobs designed in Talend Studio and retrieve a dataset directly in Talend Data Preparation through the use of the Live dataset feature.

In a Big Data context, the Streams Runner service is optionally used to trigger access to the Spark Job Server and the Hadoop Cluster, in order to import datasets from the cluster, and run preparation directly on this framework.

You can optionally use Talend Dictionary Service to add, remove or modify the semantic categories that are applied to each column in your data when opened in Talend Data Preparation.

- 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 Talend Studio and the Monitoring database is used to monitor the execution of technical processes and service calls.

The MongoDB database is used to store all your datasets and preparations, as well as the semantic types used to validate your data in the application. Nothing is saved directly on your computer.

- The **Execution server** block contains a Spark Job Server used to manage the exports that will be performed on the Hadoop cluster, and the Components Catalog.

Thanks to the Components Catalog service, you can import data stored on various types of databases and create remote datasets directly in Talend Data Preparation.

- The **Hadoop cluster** block, where preparations made on data imported from HDFS or JDBC can be processed when using Talend Data Preparation in a Big Data context.

Creating Talend Data Preparation users

Talend Administration Center allows you to define Talend Data Preparation users and assign them predefined roles.

This makes the user list accessible from Talend Data Preparation to share datasets or preparation with other listed users for example.

These users can either be related only to Talend Data Preparation, or to hybrid projects with both Data Preparation and other project types, Data Integration for example.

Creating a Data Preparation-only type of user

This type of user can access Talend Data Preparation, but does not have access to other projects.

1. Log in to Talend Administration Center.
2. In the **Users** tab of the **Menu**, click **Add**.

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

Data

Login:

First name:

Last name:

Password:


Svn login:

Svn password:


GIT login:

GIT password:


Type: ▼

Role: 

Data Preparation User: ☒

Data Preparation Role: 

Data Stewardship User: ☐

Group: 

Active: ☒

3. Enter the user's name, login (email address) and password for this account.
4. Select the **Data Preparation User** check box to set this account as a Data Preparation account
5. Set the Data Preparation user **Type** to **No Project access** as this user is not linked to any projects and will only work in Talend Data Preparation.
6. Click the pen icon next to the **Data Preparation Role** to open a dialog box where you can select from the list the check box of the Data Preparation role(s) you want to assign to the selected user.
For more information, see [User Roles](#) on page 151.
7. Click the pen icon next to the **Group** field to open a dialog box where you can select from the list the check box of the user group(s) in which you want to add the selected user.
For more information on user groups, see [Grouping users by user type](#) on page 9.
8. Click **Save** to validate the creation of the new user.

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




Creating a hybrid Data Preparation user

This type of user can work in Talend Data Preparation, as well as the project type selected from the **Type** list. For example, a hybrid Data Preparation/Data Quality user can be assigned to a **Data Quality** user group.

1. Log in to Talend Administration Center.
2. In the **Users** tab of the **Menu**, click **Add**.

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

Data

Login:	<input type="text" value="user@dataprep.com"/>
First name:	<input type="text" value="Nicolas"/>
Last name:	<input type="text" value="Talend"/>
Password:	<input type="password" value="*****"/>
Svn login:	<input type="text"/>
Svn password:	<input type="password"/>
GIT login:	<input type="text"/>
GIT password:	<input type="password"/>
Type:	<input type="text" value="Data Quality"/> ▼
Role:	<input type="text" value="Administrator/Operator"/> 
Data Preparation User:	<input checked="" type="checkbox"/>
Data Preparation Role:	<input type="text" value="Administrator/Data Steward"/> 
Data Stewardship User:	<input type="checkbox"/>
Group:	<input type="text" value="dq_group"/> 
Active:	<input checked="" type="checkbox"/>

3. Enter the user's name, login (email address) and password for this account.
4. In the **Type** field, select the type of project that the Data Preparation 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 the pen icon next to the **Role** field to open a dialog box where you can select from the list the check box of the role(s) you want to assign to the selected user.
6. Select the **Data Preparation User** check box to set this account as a Data Preparation account.
7. Click the pen icon next to the **Data Preparation Role** to open a dialog box where you can select from the list the check box of the Data Preparation role(s) you want to assign to the selected user.

For more information, see [User Roles](#) on page 151.

8. Click the pen icon next to the **Group** field to open a dialog box where you can select from the list the check box of the user group(s) in which you want to add the selected user.

For more information on user groups, see [Grouping users by user type](#) on page 9.

9. Click **Save** to validate the creation of the new user.

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

Grouping users by user type

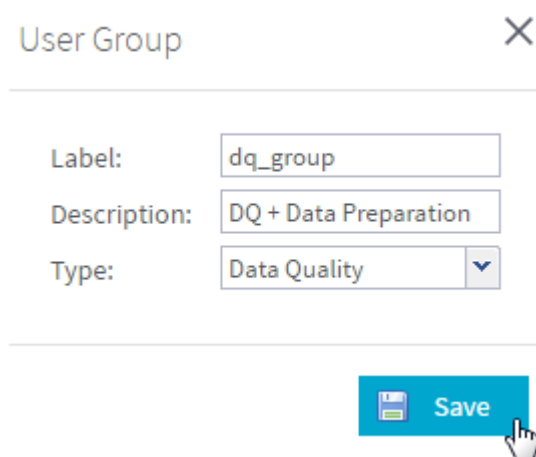
From the **User Groups** page of Talend Administration Center, you can organize existing users in groups based on their type: **Data Integration/ESB**, **Data Quality**, **Master Data Management**, **Data Preparation** or **Data Stewardship**.

Once created, these groups can be assigned to projects of the same type.

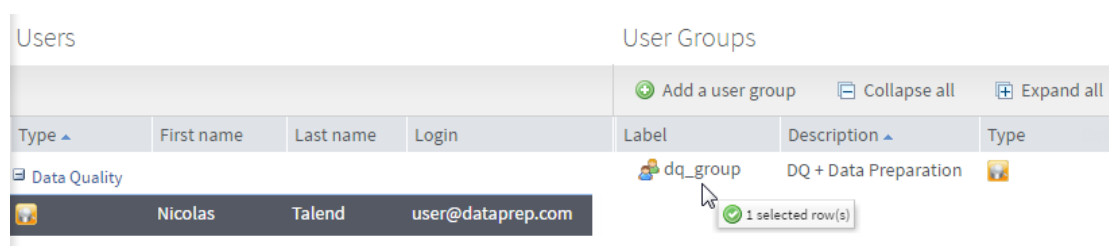
User groups allow administrators to manage large amount of users by organizing them efficiently in order to assign them easily to corresponding projects. To add users to a group, follow this procedure.

You have already created several users of the same type from the **Users** page.

1. Log in to Talend Administration Center.
2. In the **User groups** tab of the **Menu**, click **Add a user group**.
3. In the **User Group** window that opens, give a name, a type and corresponding roles, if necessary, to your user group.



4. Click **Save**.
5. From the **Users** panel of the page, select the users that you want to add in your group, then drag and drop them in the corresponding group of the **User Groups** panel.



You can select multiple users with the **Ctrl + Shift** keys.

Data Preparation-only and Data Stewardship-only users (with no related project) can only be added to Data Preparation and Data Stewardship groups respectively, whereas hybrid users can only be assigned to the group type corresponding to their own user type.

In this example, the user has the **Data Quality** type assigned, and is a **Data Preparation** user as well. As a consequence, it can be added to a **Data Quality** user group.

Your user group is created and populated with the users you have selected.

To remove a user from an existing user group, right-click the user from the **Users** panel of the **User Groups** page and click **Remove assignment**.

Talend Data Preparation start and stop sequences

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

Zookeeper and Kafka are necessary only if you are using Talend Dictionary Service to enrich semantic types. The availability of Talend Dictionary Service depends on your subscription level.

Streams Runner and Spark Job Server are necessary only if you are using Talend Data Preparation for Big Data.

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

1. Zookeeper
2. Kafka
3. MongoDB
4. Talend Administration Center
5. Talend Identity and Access Management Service
6. Talend Dictionary Service
7. Components Catalog
8. Streams Runner
9. Spark Job Server
10. Talend Data Preparation

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

Talend Data Preparation log files location

Talend Data Preparation logs allows you to analyze and debug the activity of Talend Data Preparation.

By default, Talend Data Preparation will log in two different places, namely the console, and in a log file.

The location of this log file depends on the version of Talend Data Preparation that you are using:

- `<Data_Preparation_Path>/data/logs/app.log` for Talend Data Preparation.

- AppData/Roaming/Talend/dataprep/logs/app.log for Talend Data Preparation Free Desktop on Windows.
- Library/Application Support/Talend/dataprep/logs/app.log for Talend Data Preparation Free Desktop on MacOS.

These locations can be configured by editing the `logging.file` property of the `application.properties` file.

Configuring the Talend Data Preparation log files

To configure the settings of your log files, edit the `<Data_Preparation_Path>/config/application.properties` file.

Specify patterns for your logs by using `logging.pattern.<file|console>` properties. These properties use a logback style pattern. The default logging pattern is `logging.pattern.level=%5p [user %X{user}]`, which displays the log level information, and the user's ID.

For more information, see the documentation on [Logback pattern layouts](#).

Let's take the example of the following pattern, that will apply to the log files but not the console:

```
logging.pattern.file=%d{yyyy-MM-dd HH:mm:ss.SSS} [%thread] %-5level
%logger{15} - %msg %n
```

The output in the log file will look like this:

```
2017-02-21 11:07:03.741 [main] WARN c.l.TrivialMain - a warning message 0
2017-02-21 11:07:03.741 [main] DEBUG c.l.TrivialMain - hello world number1
2017-02-21 11:07:03.741 [main] DEBUG c.l.TrivialMain - hello world number2
2017-02-21 11:07:03.741 [main] INFO c.l.TrivialMain - hello world number3
2017-02-21 11:07:03.741 [main] DEBUG c.l.TrivialMain - hello world number4
2017-02-21 11:07:03.741 [main] WARN c.l.TrivialMain - a warning message 5
2017-02-21 11:07:03.741 [main] ERROR c.l.TrivialMain - Finish off with fireworks
```

Using a logback configuration for the Talend Data Preparation log files

If you want to enable a more advanced logging configuration, you will need to specify a `logback.xml` file in the `application.properties` file.

This `logback.xml` file containing your configuration is specified using the `logging.config=<path-to-config-file>` property.

Example of a logback configuration to set the logging to the DEBUG level:

```
<configuration>

  <appender name="STDOUT" class="ch.qos.logback.core.ConsoleAppender">
    <!-- encoders are assigned the type
         ch.qos.logback.classic.encoder.PatternLayoutEncoder by default -->
    <encoder>
```

```

    <pattern>%d{HH:mm:ss.SSS} [%thread] %-5level %logger{36} - %msg%n</pattern>
  </encoder>
</appender>

<root level="debug">
  <appender-ref ref="STDOUT" />
</root>
</configuration>

```

For more information, see the documentation on the [Logback file configuration](#) and the [Spring logging system](#).

Backing up your Talend Data Preparation instance

Backing up Talend Data Preparation 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 Preparation instance, you need to backup the MongoDB, the folders containing your data, the configuration files and the logs.

This procedure backs up MongoDB with mongodump, but you can use different 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 instance.
2. Stop your Talend Data Preparation instance.
3. To backup your MongoDB, open a command prompt window and execute the following command:


```

mongodump -h <source_mongodb_host>:<source_mongo_port>
-d <source_database> -u <source_mongodb_user> -p
<source_mongodb_password> -o <dump_output>

```
4. To back up your datasets, make a copy of the folder specified in the `dataset.content.store.file.location` property of the `<Data_Preparation_Path>/config/application.properties` file, and save it to your preferred location.
5. To back up any custom configuration that you might have, make a copy of the `<Data_Preparation_Path>/config` folder and save it to your preferred location.
6. To back up your log files, make a copy of the `<install_folder>\dataprep\data\logs` folder and save it to your preferred location.
7. 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 using Talend Dictionary Service, make a copy of the folder specified in the `dataquality.indexes.file.location=<preferred_location>/org.talend.dataquality.semantic` property of the `<Data_Preparation_Path>/config/application.properties` file.
 - b) Back up the Talend Dictionary Service server. For further information, see [Talend Dictionary Service back up and recovery](#) on page 13.

You now have a backup of your Talend Data Preparation instance.

Restoring your Talend Data Preparation instance

After backing up your Talend Data Preparation instance, you can restore your database and configuration at any time.

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

1. Start your MongoDB instance.
2. Stop your Talend Data Preparation instance.
3. To restore your MongoDB, open a command prompt window and execute the following command:

```
mongorestore -h <source_mongodb_host>:<source_mongo_port>
-d<source_database> -u <source_mongodb_user> -p
<source_mongodb_password> dump\
```
4. To restore your datasets, replace the content of the folder specified in the `dataset.content.store.file.location` property of the `<Data_Preparation_Path>/config/application.properties` file, with your backed up copy.
5. To restore your custom configuration, replace the content of the `<Data_Preparation_Path>/config` folder with your backed up copy.
6. To restore your log files, replace the content of the `<install_folder>\dataprep\data\logs` folder with your backed up copy.
7. 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 the folder specified in the `dataquality.indexes.file.location=<preferred_location>/org.talend.dataquality.semantic` property of the `<Data_Preparation_Path>/config/application.properties` file 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 13.

Talend Data Preparation now uses the source files, as well as the configuration, that you had previously saved.

Talend Dictionary Service back up and recovery

The availability of the Dictionary service depends on the license you have.

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.

Talend Dictionary Service stores all the predefined semantic types used in Talend Data Preparation. It also stores all the custom types created by users and all the modifications done on existing ones.

To back up a Talend 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 `<pathToTalendInstaller>/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. Launch MongoDB.

3. Open a command prompt window and execute the following commands to backup the current Talend Dictionary Service instance, including the logs, to a dump folder:
on Windows:

- `<pathToTheInstallationFolder>/dq_dict/database`
- `semantic-dictionary-export.bat`

on Linux:

```
<pathToTheInstallationFolder>/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 Talend Dictionary Service instance is completed and the file is saved at
`<pathToTalendInstaller>/dq_dict/database/dump`.

You will also need to back up the changes made to the predefined semantic types using Talend Dictionary Service and stored in the folder specified in the `<Data_Preparation_Path>/config/application.properties` file. For further information, see [Backing up your Talend Data Preparation instance](#) on page 12.

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 following command to delete the database in the current instance of Talend Dictionary Service:

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

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

```
<pathToTheInstallationFolder>/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 got 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 Talend Dictionary Service instance is recovered.

Defining the maximum number of column for a dataset

To prevent the import of datasets with an abnormally large numbers of columns, a property can be added to the Talend Data Preparation configuration file to set a maximum number of columns to your datasets.

In other words, you cannot import a dataset with a number of columns exceeding the value set in the `dataset.import.xls.size.column.max` property of the `<Data_Preparation_Path>/config/application.properties` file. By default, this value is set to 1000.

This limit applies to all dataset types.

To change this maximum number, add the `dataset.import.xls.size.column.max` property with the desired value.

For example: `dataset.import.xls.size.column.max=50`

F.A.Q.

File formats supported by Talend Data Preparation

In Talend Data Preparation, you can import different types of file to use as source data for your datasets.

From local files

You can import the following file types to use as datasets:

- `.xls` or `.xlsx`
- `.csv`

For more information, see [Adding a dataset from a local file](#) on page 86.

From a Talend Job

In addition to the previous file types, you have the possibility to use datasets created directly from a Talend Job in Talend Studio if you are a subscription user.

You can do that by using the `tDatasetOutput` component as output for your Job in Talend Studio.

Then you can either:

- Run the Job directly in Talend Studio. For more information, see [Creating a dataset from a Talend Job](#) on page 106.
- Use the live dataset feature to run it via Talend Administration Center and access the data directly in Talend Data Preparation. For more information, see [Creating a dataset based on an on-demand Job execution](#) on page 112.

From a database

Talend Data Preparation is able to connect to various databases and use them as source to create a new dataset. The data is still stored in your database, and only a sample is retrieved on-demand.

For more information, see [Adding a dataset from a database](#) on page 88.

From HDFS

You can access data that is stored on a Hadoop file system (HDFS), and import it in the form of a dataset, directly in the Talend Data Preparation interface. You can then export the prepared data back to the cluster, or export it as a local file.

For more information, see [Adding a dataset from HDFS](#) on page 93.

From Salesforce

You can access data that is stored on Salesforce, and import it in the form of a dataset, directly in the Talend Data Preparation interface. You can then export the data as a local file.

For more information, see [Adding a dataset from Salesforce](#) on page 97.

From Amazon S3

You can access data that is stored on Amazon S3, and import it in the form of a dataset, directly in the Talend Data Preparation interface. You can then export the data as a local file, export it on a Hadoop cluster, or export it back directly to Amazon S3.

For more information, see [Adding a dataset from Amazon S3](#) on page 101.

Data storage location

According to the version of Talend Data Preparation that you are using, your data is stored in different locations.

Talend Data Preparation

If you are a subscription user, nothing is saved directly on your computer. All the data from your datasets and preparations is stored remotely on the Talend Data Preparation server.

When using the live dataset feature, or preparing data stored on a database or a Hadoop file system (HDFS,) the sample data is cached on the Talend Data Preparation server for one hour.

Talend Data Preparation Free Desktop

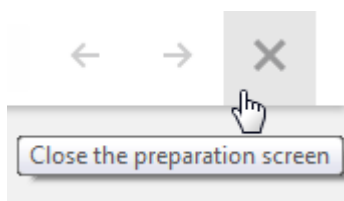
Talend Data Preparation Free Desktop is meant to be able work locally on your computer, without the need of an internet connection. Therefore, when using a dataset from a local file such as a CSV or Excel file, the data is copied locally, in one of the following folders depending on your operating system:

- Windows: C:\Users\\AppData\Roaming\Talend\dataprep\store
- OS X: /Users/<your_user_name>/Library/Application Support/Talend/dataprep/store

Lock mechanism for shared datasets and preparations

The sharing feature in Talend Data Preparation includes a locking system to prevent users from overwriting the work of the others. In other words, only one person is able to access a shared preparation or dataset at a time.

But if one of the authorized user did not exit the dataset or preparation properly, by clicking the **Close the preparation screen** button, it will still be considered locked for the other users, even if no one is currently working on it.



You have to wait 10 minutes in order to have access to the preparation or dataset again. Before exiting Talend Data Preparation, remember to close the dataset or the preparation using the corresponding button, instead of just closing the Talend Data Preparation window in your Web browser.

Managing datasets

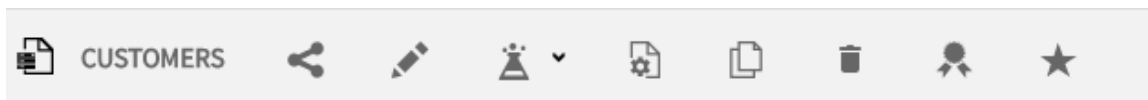
Actions on datasets

When pointing your mouse over your datasets, several actions are available to manage or sort them.








Dataset based on a local file:





Dataset based on a database or located on a HDFS cluster:



The table below describes the different actions that you can perform on your datasets.

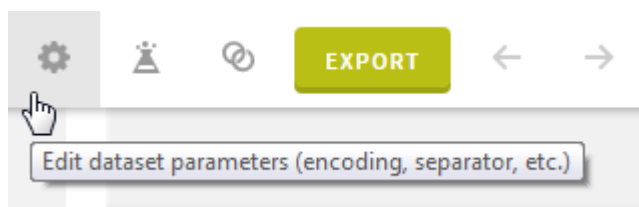
Icon	Action	Description
	Rename your dataset	After typing the new name of your dataset, click the green tick icon to validate.
	Share your dataset	Use this button to make your datasets accessible to the other members of your organization. They will be able to open and add preparations on your datasets.
	Apply an existing preparation on your dataset	When you add a dataset with a schema that is identical to another dataset you already have, you can apply an existing preparation to it.
	Overwrite your dataset with another file	If the local file you used as a dataset has changed since you created the dataset, you can refresh this dataset with the latest data.
	Edit your dataset	This button opens a menu where you can modify the parameters that you entered to access a dataset located on a database or HDFS cluster. You can also change the name of the dataset.
	Copy your dataset	You can duplicate a dataset if you need more than one copy of the same dataset. You will find the new copy at the top of the Datasets list.
	Delete your dataset	A confirmation dialog opens before deleting the dataset.

Icon	Action	Description
	Certify your dataset	When doing collaborative work, user with their role set as Administrator can certify a dataset to indicate that this dataset has been validated.
	Add a dataset to your favorites	Adding a dataset to your favorites makes it easier for you to find when creating a new preparation based on your datasets.

Changing the encoding

When adding a new dataset, Talend Data Preparation automatically guesses the encoding to be used.

1. Open the dataset for which you want to change the encoding.
2. Click the cog icon to open the **Datasets parameters** dialog box.



3. In the **Encoding** list, choose the encoding that corresponds to your dataset.

Dataset parameters

Name	customers	Encoding	ISO-8859-1 ▼
Size	1432 rows	Separator	; ▼

CONFIRM

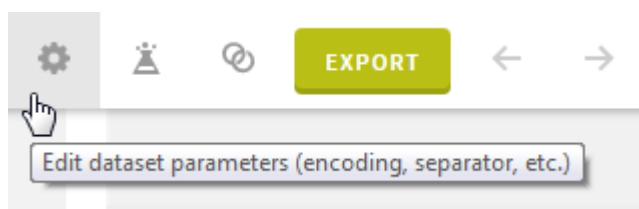
4. Click the **Confirm** button.

The dataset reloads and the new encoding is applied.

Changing the separator

When adding a new dataset from a local `.csv` file, Talend Data Preparation automatically guesses the separator (or delimiter) to be used, but you can change it.

1. Open the dataset for which you want to change the separator.
2. Click the cog icon to open the **Dataset parameters** dialog box.



3. In the **Separator** drop-down list, choose the separator that corresponds to your dataset.

Dataset parameters

Name	customers	Encoding	ISO-8859-1
Size	1432 rows	Separator	;

CONFIRM

4. Click the **Confirm** button.

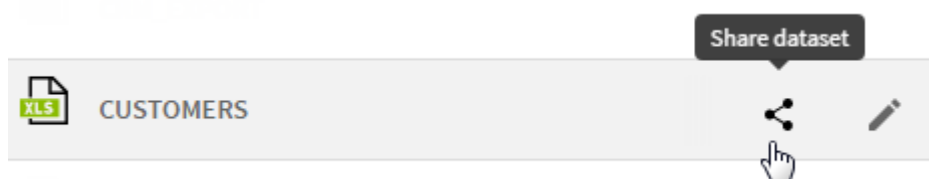
The dataset reloads and the new separator is applied.

Sharing a dataset

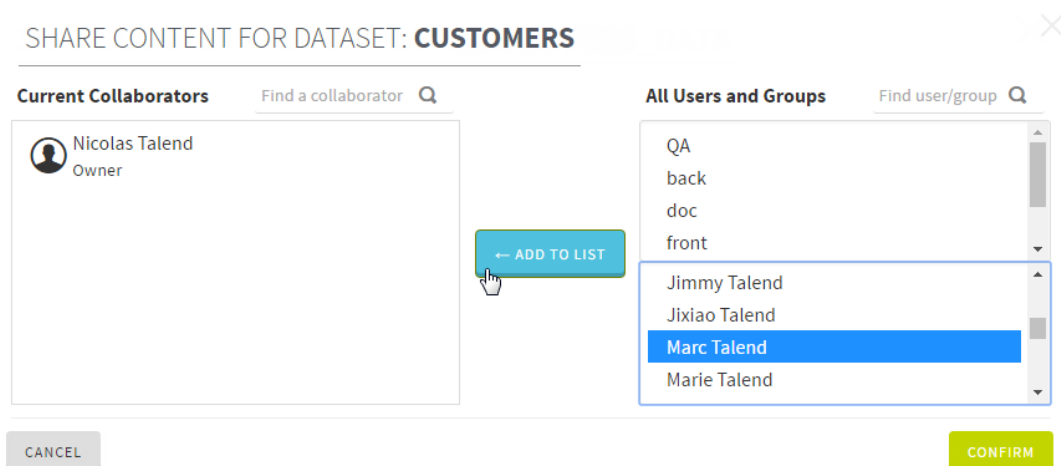
To make your datasets accessible to the other members of your organization, you can use the share feature.

Other users or groups of users will be able to open and add preparations on your datasets.

1. Click **Datasets** to open the list of datasets.
2. Point your mouse over the dataset you want to share in order to display the available options.
3. Click the share icon to open the **Share Content for Dataset** window.



4. Browse the **All Users and Groups** list or use the **Find user/group** search bar to select a user or group.
5. Click a user or group and click **Add to List** to add them to the list of contributors.



6. You can repeat the last step to add more contributors.

7. Click **Confirm**.

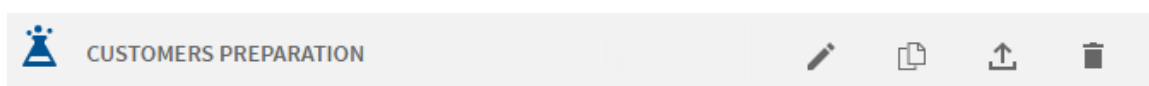
The selected users or groups have been added to the list of contributors and they can now see your dataset in the **Datasets** view.

Managing preparations



Actions on preparations



When pointing your mouse over your preparations or preparation folders, several actions are available to manage or sort them.

Actions on preparations

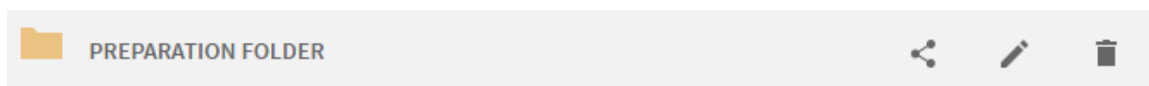


The table below describes the different actions that you can perform on your preparations.




Icon	Action	Description
	Rename preparation	After typing the new name of your preparation, click the green tick icon to validate or the grey cross icon to cancel.
	Copy or move preparation	You can duplicate or move your preparations to the selected location in the folder structure.

Icon	Action	Description
		You can enter a new name or keep the current one.
	Export preparation	Use this button to export a preparation to a json file, and then promote it to another environment.
	Delete preparation	A confirmation dialog opens before deleting the preparation.

Actions on preparation folders



The table below describes the different actions that you can perform on your preparation folders.

Icon	Action	Description
	Rename preparation folder	After typing the new name of your folder, click the green tick icon to validate or the grey cross icon to cancel.
	Share preparation folder	Use this button to make your preparation folders accessible to the other members of your organization. They will be able to open modify the preparations they contain.
	Delete preparation folder	You can only delete empty folders.

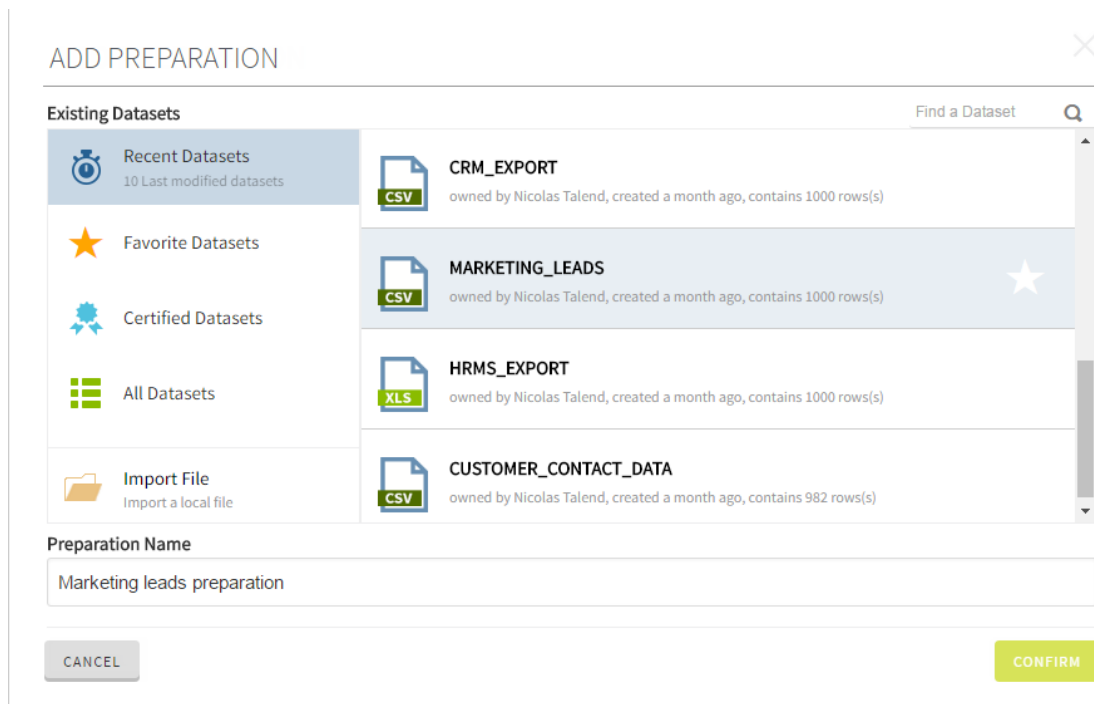
Adding a preparation

Add a preparation to start preparing and cleansing your data.

You can create a preparation from a dataset already available in Talend Data Preparation or one of your local files. When you add a preparation with the corresponding button, it will be created in the folder in which you are currently working. Furthermore, your preparation will automatically be saved in the preparations list, and all the changes you make are also saved automatically.

1. Click **Preparations** to open the list of preparations.

2. Click the **Add Preparation** button.



3. In the **Preparation Name** field, enter the name you want to give your preparation.

4. To choose which dataset to use for your preparation, you can:

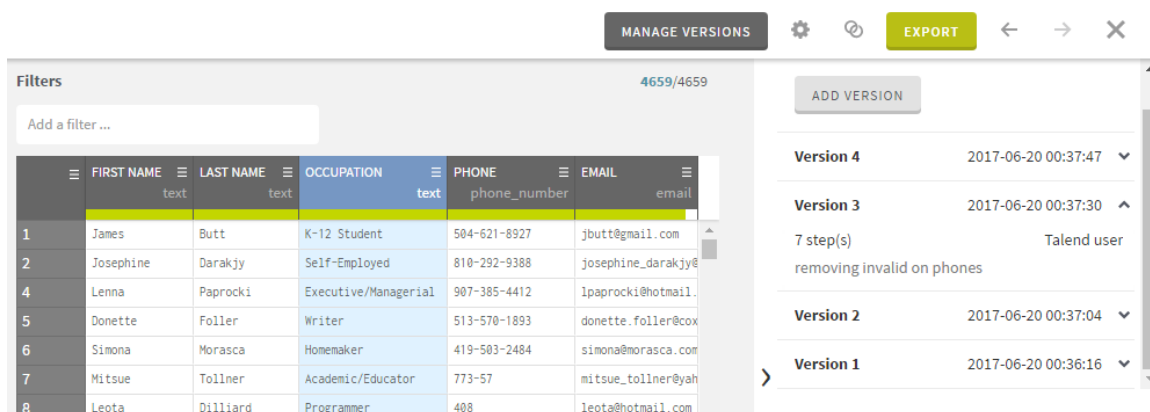
- Either select a dataset from the a list of already existing datasets and click **Confirm**.
- Or click **Import File** to directly add a new one from a local file.

Your dataset opens with an empty recipe, and you can start adding preparation steps. All your changes are automatically saved.

Preparation versioning

When working on your data, you can decide to capture the state of your preparation by creating a version.

Creating a version can be done at any moment, even when no steps have been applied yet. It allows you to freeze a preparation in a given state, with a timestamp and a short description.



Use the **Manage versions** button to create a new version of your preparation, or consult previously created version in read-only mode. Each version can be individually exported.

Adding versions to your preparation is a good way to see the differences that have been made to the preparation over time, but mostly to ensure that it is always the same state of a preparation that is used in Talend Jobs, even if the preparation is still being worked on. Versions can be used in Data Integration as well as Big Data Jobs.

Preparation versions are propagated when sharing or moving a preparation across your folder structure, but not when you copy it or apply it to a new dataset.

Creating preparation versions

In the following example, you will perform a few preparation steps on your data, create versions at two different moments, and see how you can switch between your versions, as well as switch back to the current state of your preparation.

The dataset used here contains customer data such as their names, occupation, phone number and email address, but that requires some cleansing. Formatting inconsistencies can be found in the columns containing the customers names, such as leading or trailing whitespaces, and inconsistent case. In addition, various phone and email entries are invalid.

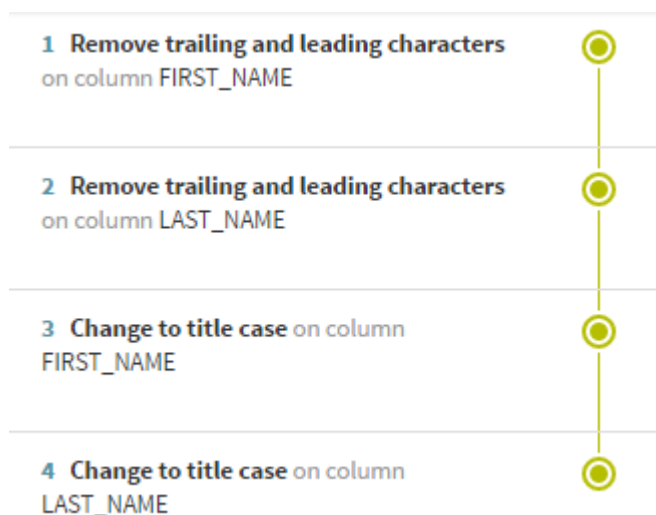
	FIRST_NAME text	LAST_NAME text	OCCUPATION text	PHONE phone_number	EMAIL email
1	James	Butt	K-12 Student	504-621-8927	jbutt@gmail.com
2	Josephine	Darakjy	Self-Employed	810-292-9388	josephine_darakjy@da
3	ART	Venere	Scientist	856-636-	art@venere
4	Lenna	Paprocki	Executive/Managerial	907-385-4412	lpaprocki@hotmail.co
5	Donette	Foller	Writer	513-570-1893	donette.foller@cox.n
6	Simona	Morasca	Homemaker	419-503-2484	simona@morasca.com
7	Mitsue	Tollner	Academic/Educator	773-57	mitsue_tollner@yahoo
8	Leota	dilliard	Programmer	408	leota@hotmail.com
9	Sage	Wieser	Technical/Engineer	605-414-2147	sage_wieser@cox.net

As you progress in your preparation, you are going to create two versions, that reflect the state of your preparation at two different times.

1. Click the header of the **FIRST_NAME** column, and while pressing the **Ctrl** key, click the header of the **LAST_NAME** column.

The content of the two columns is now selected.

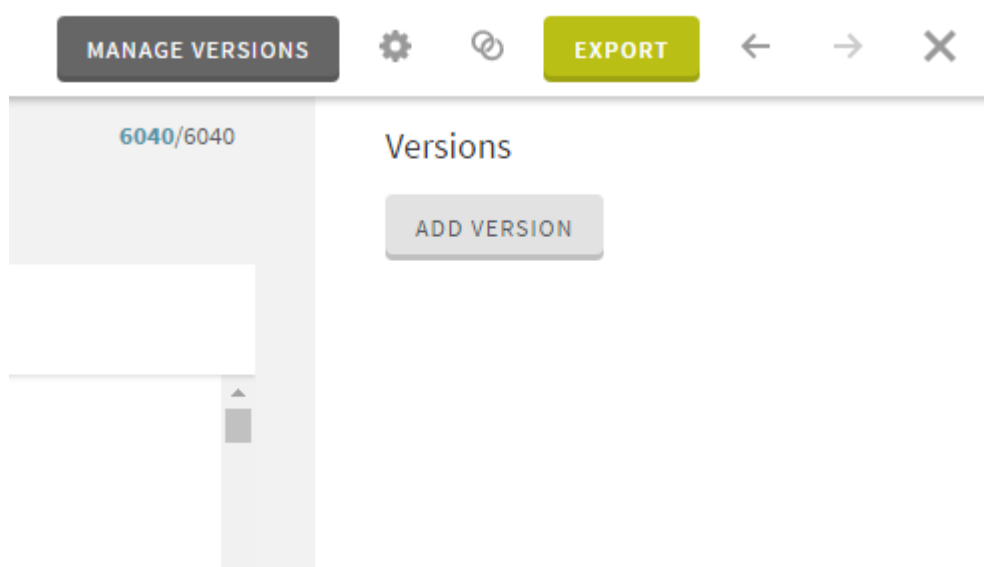
2. Apply the **Remove trailing and leading characters** and the **Change to title case** functions to remove whitespaces and harmonize the case.



Removing those formatting errors marks the first big step in your preparation, and you are going to create a version to track these changes.

3. Click the **Manage versions** button located in the header bar.

The **Functions panel** is replaced with the **Versions** panel. This panel is empty since no versions exist for this preparation yet.



Adding new versions via the **Manage versions** button is only available to Talend Data Preparation user with administrator rights. Other users are only able to consult existing version in read-only mode.

4. Click the **Add version** button.
5. Enter a quick description of the version in the corresponding field, *Fixing formatting errors in names* in this example, and click **Submit**.

The version is now listed in the **Versions** panel with a timestamp, and the description you added before.

6. Click the version to access it in read-only mode.
You can apply filters and browse the data, but you cannot apply functions on it.
7. To leave the read-only mode and resume preparing your data, click the **Switch to current state** button located in the header bar.
You are now back to the edit mode.
8. To cleanse the remaining invalid entries from the **PHONE** and **EMAIL** columns, click the menu icon on the top left corner of the grid, and select **Display rows with invalid or empty values**.
9. From the **Functions panel**, select the **Delete these filtered rows** functions.

All the invalid values have been removed from your dataset, and you are going to create another version to capture this state.

10. Repeat steps 3 to 5 to create a new version, but this time, enter **Removing all invalid values** as description.

Your two versions are now listed in the **Versions** panel and can be accessed in read-only mode.

Versions		
ADD VERSION		
Version 2	2017-06-19 15:21:40	▼
Version 1	2017-06-19 14:26:23	▼

You have created two versions of your preparation, in order to capture its state at two different steps of the cleansing process. You can choose to export one of these versions, use it in a Talend Job, or continue to edit the current state of your preparation.

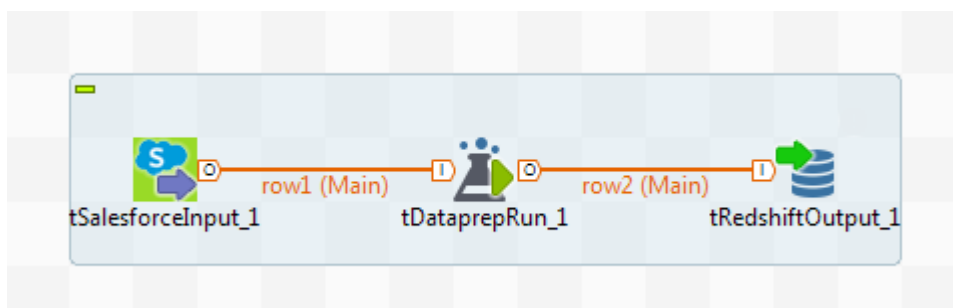
Using a version in a Talend Job

Preparation versions can be used in data integration or Big Data Jobs in Talend Studio.

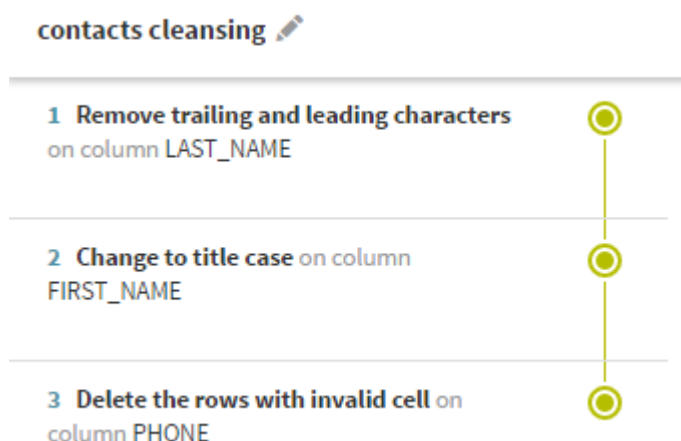
In Talend Studio, the tDataprepRun component allows you to reuse a preparation, or any of its versions, and apply it on data with the same model.

You still have the possibility to use a preparation in its current state, but using a specific version can ensure that it is always the same state of a preparation that is used in your Jobs, even if the preparation is still being worked on, thus providing more consistency.

The following example will illustrate a Job that applies an existing preparation version on a Salesforce input, and outputs it to a Redshift database.



This preparation was made on a dataset containing basic customer information such as names, phone numbers and email addresses. A few steps have been applied to remove formatting errors in the name entries, and to delete invalid values from the phone numbers.



Two versions have been created during the preparation: one after the first two steps, and another one after the third step.

Versions		
<button>ADD VERSION</button>		
Version 2	2017-06-19 15:21:40	▼
Version 1	2017-06-19 14:26:23	▼

- You have created a preparation with at least one version in Talend Data Preparation. In this case the existing preparation is called **contacts cleansing**.
 - The data imported from salesforce must have the same schema as the dataset used to create the preparation in the first place.
1. In Talend Studio, create a new Standard or Spark Job.
 2. In the design workspace of Talend Studio, add a **tSalesforceInput**, a **tDataprepRun**, a **tRedshiftOutput**, and link them together using two **Row > Main** links.
 3. Select the **tSalesforceInput** component and click the **Component** tab to define its basic settings. Make sure that the schema of the **tSalesforceInput** component matches the schema expected by the **tDataprepRun** component.
 4. Select the **tDataprepRun** component and click the **Component** tab to define its basic settings.

tDataprepRun_1

Basic settings

Advanced settings

Dynamic settings

View

Documentation

Validation Rules

Data Preparation Connection

URL "http://localhost:9999"

Username "user@dataprep.com" Password *****

Configuration

Preparation "contacts cleansing" Choose an existing preparation Or create a new one

Version 1 Choose a Version

Fetch Schema Schema Built-In Edit schema Sync columns

5. Enter your Talend Data Preparation connection information.
6. Click **Choose an existing preparation** to display a list of the preparations available in Talend Data Preparation.

Select an existing preparation

Name	Author	Last Modification
<input checked="" type="checkbox"/> contacts cleansing	User3 User3	21/06/17 13:44
<input type="checkbox"/> customers Preparation	User3 User3	20/06/17 18:36
<input type="checkbox"/> uk_customers	User3 User3	20/06/17 00:37
<input type="checkbox"/> datapreun_spark_test	User3 User3	02/06/17 17:25

OK Cancel

7. Select the checkbox in front of **contacts cleansing**, that contains the preparation version that you want to apply, and click **OK**.
8. Click **choose a version** to select from the list of available versions for your preparation. In this case, select version **1**.

Set the version

Version	Author	Creation Date	Number of steps
<input type="checkbox"/> Current state			
<input type="checkbox"/> 2	User3 User3	19/06/17 15:21	3
<input checked="" type="checkbox"/> 1	User3 User3	19/06/17 14:26	2

OK Cancel

By default, the Job uses the **current state** of the selected preparation. Using the **current state** instead of a fixed version means that in the context of collaborative work, someone possibly made changes, that you are unaware of, on the preparation. As a consequence you cannot know exactly what the outcome of your Job will be. This is why it is safer to use a version in your Jobs.

9. Click **Fetch Schema** to retrieve the schema of **contacts cleansing**.
10. Select the **tRedshiftOutput** component and click the **Component** tab to define its basic settings.
11. Save your Job and press **F6** to run it.

All the preparation steps included in the version of the preparation have been applied to your data, directly in the flow of your Job.

Promoting a preparation across environments

The best practice when using Talend Data Preparation is to set up one instance for each environment of your production chain.

To promote a preparation from one environment to the other, you have to export it from the source environment, and then import it back to your target environment.

For the import to work, a dataset with the same name and schema as the one which the export was based on must exist on the target environment.

Let's take the example of a simple preparation, used to remove invalid values and formatting errors from an Excel file containing customers data. This preparation is saved as `customers preparation`.

customers preparation

- 1 Remove trailing and leading characters on column LAST_NAME
- 2 Delete the rows with invalid cell on column ZIP
- 3 Change to title case on column FIRST_NAME
- 4 Replace the cells that match on column STATE
- 5 Delete the rows with invalid cell on column PHONE
- 6 Change date format on column SUBDATE

Filters

Add a filter ...

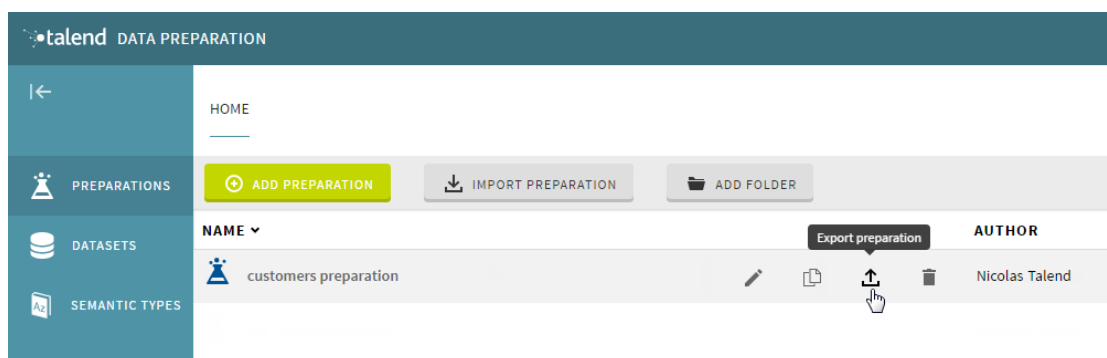
	First_Name	Last_Name	Gender	Age	Occupation
	first_name		gender		
1	James	Butt	F	Under 18	K-12 Student
2	Josephine	Darakjy	M	56+	Self-Employed
4	Lenna	Paprocki	M	45-49	Executive/Managerial
5	Donette	Foller	M	25-34	Writer
6	Simona	Morasca	F	50-55	Homemaker
9	Sage	Wieser	M	25-34	Technical/Engineer
13	Kiley	Caldarera	M	45-49	Academic/Educator
14	Graciela	Ruta	M	35-44	Other
15	Cammy	Albares	M	25-34	Executive/Managerial
16	Mattie	Poquette	F	35-44	Other
17	Meaghan	Garufi	M	50-55	Academic/Educator

The following procedure will detail how to export this simple preparation into a file that can be reimported to a new environment.

Exporting a preparation from the source environment

Now that your preparation has been finalized on your source environment, it is ready to be exported and sent to the target environment.

1. Log in to your Talend Data Preparation source environment.
2. In the left panel menu, click **Preparations** to open the list of existing preparations.
3. Point your mouse over **customers preparation** to display the available actions for this preparation.
4. Click the **Export preparation** button.



The preparation is exported as a `.json` file and the download automatically starts. Save it to your preferred location and rename it `customers_preparation.json` for example.

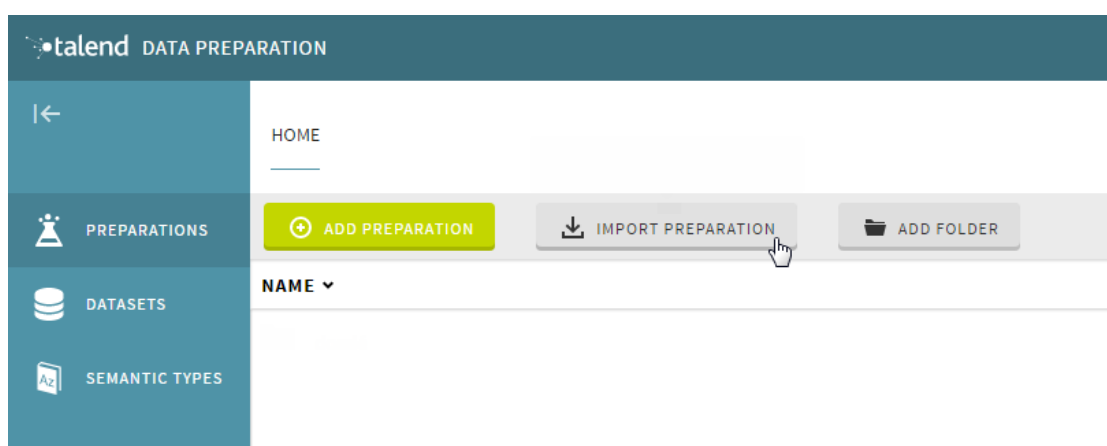
Importing a preparation to the target environment

You will now access your target environment and import the previously exported preparation.

A dataset with the same name and schema as the one which the export was based on must be present on the target environment.

The reason for this is that when a preparation is exported as a `.json` file, the name and schema of the source dataset are exported, in addition to the preparation steps that were applied. The target environment must contain this information on the source dataset, in other words a dataset with the same name and schema, or an error will occur during the import.

1. Log into your Talend Data Preparation target environment.
2. In the left panel menu, click **Preparations** to display the available actions related to preparations.
3. Click the **Import preparation** button.



4. In the file explorer window, select the `customers_preparation.json` file previously exported.
A loading bar is displayed during the import process. If a preparation with the same name already exists, you can either overwrite it, or import your preparation with a different name.

The customers preparation is added to the list of preparations in the **Preparations** view of the homepage. It can now be safely used as a production-ready item, in a Talend Job for example, to easily clean data with the same model.

Saving a preparation

In order not to lose any changes you make to your preparation, you should save it.

1. On the upper left part of the screen, enter a name for your preparation.



2. Click the green tick icon to validate your choice.

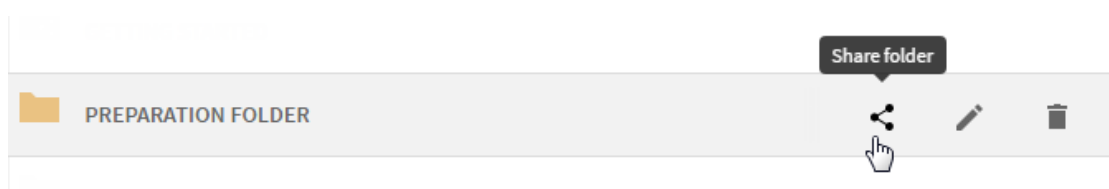
Your preparation is saved. Every modification you make from now on will automatically be saved in the preparation you created.

Sharing a preparation

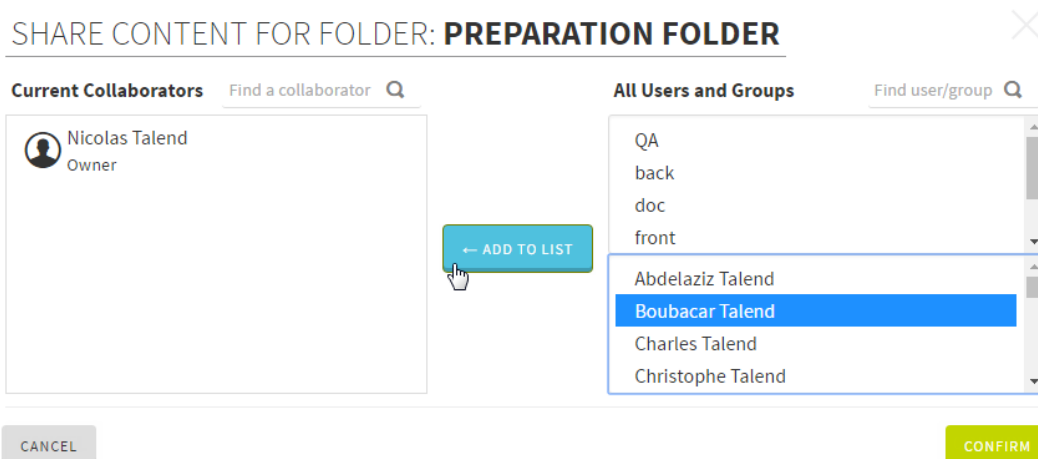
To make your preparations accessible to the other members of your organization, you can share the folder in which they are located.

Other users or groups of users will be able to open and edit your preparations.

1. Click **Preparations** to open the list of preparations and folders containing preparations.
2. Point your mouse over the folder you want to share in order to display the available options.
3. Click the share icon to open the **Share content for folder** window.



4. Browse the **All Users and Groups** list or use the **Find user/group** search bar to select a user or group.
5. Click a user or group and click **Add to List** to add them to the list of collaborators.



6. Repeat the last step to add more contributors if necessary.

7. Click **Confirm**.

The selected users or groups have been added to the list of contributors and they can now see your folder in the **Preparations** view. They will be able to edit all the preparations contained in the shared folder, enabling collaborative work. To prevent users from overwriting the work of the others, only one person is able to access the preparation at a time.

What has been shared is the preparation, and not the dataset it is based upon.

In the **Preparations** view, you can distinguish if a folder is shared or not according to the following visual code:

- A folder you own and has not been shared.



- A folder that has been shared with you.



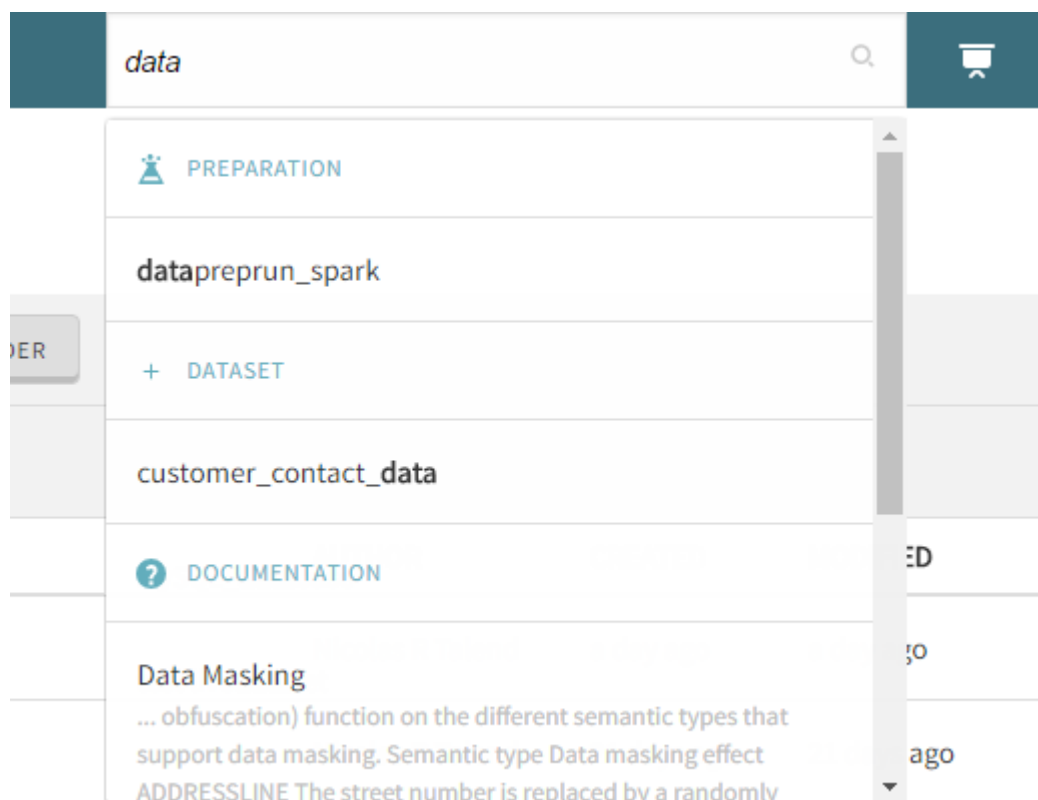
- A folder you own and has been shared.



Using the search bar

Instead of navigating through your folders, you can save time and use the search bar to search for datasets, preparations, documentation or semantic type, and open them directly.

1. In the search bar located at the top of the page, enter the first letters of the dataset, preparation, semantic type or topic that you want to open.



A list of results opens, where the text that matches your search appears in bold.

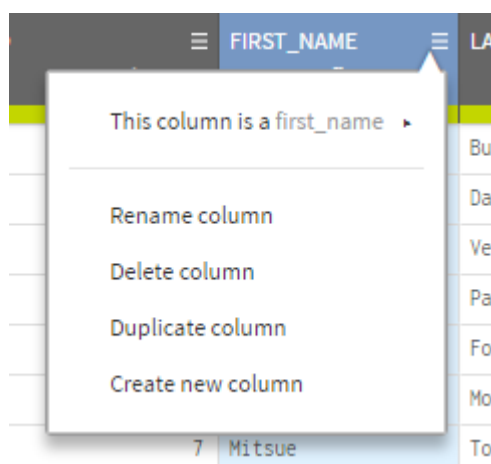
2. In the list of results, click the dataset, preparation, semantic type, or topic to open.

The item you chose directly opens.

Discovering the data

Actions on preparation columns

In a column header, click the menu icon next to the column name to display the available options.



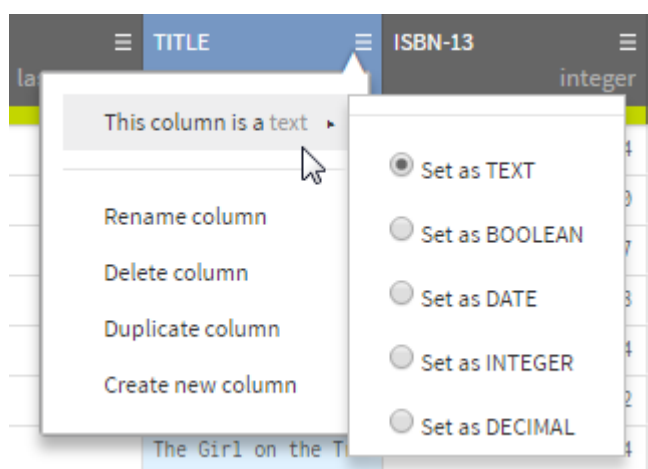
The table below describes the different actions that you can perform on columns.

Action	Description
Create column	Creates a new column to include additional data in your dataset if needed. Specify how to fill the content of the new column, either by copying an existing column or leaving it blank.
Delete column	Allows you to remove unnecessary columns.
Duplicate column	Creates a copy of a given column.
Rename column	You can rename a column with a meaningful name if you want to identify them more easily. It is also possible to rename a column by double clicking its name.
This column is a...	Allows you to change a column data type, if the value is not the desired one.

Enriching the semantic types libraries through the UI

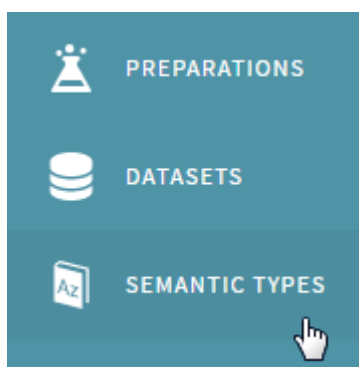
When you add a dataset, Talend Data Preparation automatically suggests one of the supported semantic types for each column.

If the semantic type proposed by Talend Data Preparation for one column is not the desired one, you can manually change it by clicking the white arrow in the column header.



This allows you to choose among the list of semantic types present in Talend Data Preparation by default. See [Predefined Semantic Types](#) on page 146 for more information. You can go further by creating your own semantic types, as well as updating or deleting the existing ones, so that Talend Data Preparation speaks your business language.

The semantic types modifications are made directly in the Talend Data Preparation interface, via the **Semantic types** tab of the left menu.



All the changes are stored using Talend Dictionary Service and are propagated across various Talend products.

The availability of Talend Dictionary Service depends on the license you have.

In Talend Dictionary Service, the semantic types are divided into three main categories:

- The **DICT** type, based on an open or closed list of values.
- The **REGEX** type that compares your data against a preselected regular expression.
- The **COMPOUND** type, under which you can group several existing types.

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

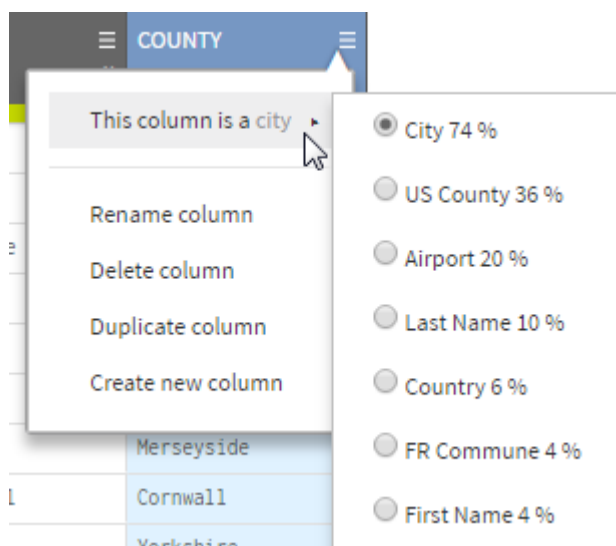
- Talend Dictionary Service is installed and running.
- Talend Administration Center is installed and running.
- Your Talend Administration Center user type is either **Master Data Management** or **Data Management**.
- Your Talend Administration Center user role is set to **Designer** or **Operation manager**.
- The **Data Preparation User** check box is selected for your user in Talend Administration Center with any of the three possible roles set in the **Data Preparation Role** field.
- In the `<install_folder>\dataprep\config\application.properties` file, the `dataquality.semantic.update.enable` property is set as `true`.

Adding a new dictionary-based semantic type through the UI

You can create a semantic type based on a closed dictionary in the **Semantic types** menu, so that it is added to the list of recognized data types.

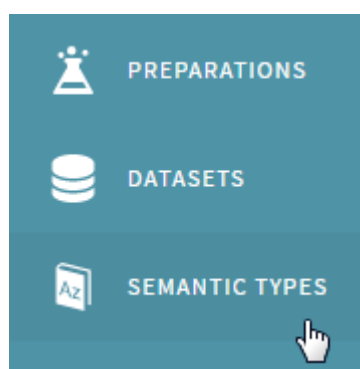
In Talend Data Preparation, not every type of data can currently be matched with one of the predefined semantic types. The counties of United Kingdom for example, are currently not recognized as such.

Let's say that you work for a British company, with customers only residing in the United Kingdom. In this example, you need to clean some customer data, such as their names, email address, or the county they live in. The semantic type for the column containing the counties data will be set by default to `city`. Some of the data may actually match names of cities, but you want to add a semantic type that is more specific to your data: `UK_counties` semantic type in this case.



You will create this new semantic type in the dedicated menu, and it will be instantly available in your preparation, so that your data can be matched with a proper type.

1. Click the **Semantic types** tab of the left menu.



The list of all the semantic types present by default in Talend Data Preparation opens. For the complete list, see [Predefined Semantic Types](#) on page 146.

talend DATA PREPARATION			
<div>←</div> <div>PREPARATIONS</div> <div>DATASETS</div> <div>SEMANTIC TYPES</div>	+ ADD SEMANTIC TYPE		
	Display: Sort by: Name Ascending		
	NAME	DESCRIPTION	TYPE
	Airport	Airport name	Dictionary
	Airport Code	Airport name	Dictionary
	Amex Card	American Express card	Regular expression

2. Click the **Add semantic type** button.

The semantic type creation form opens.

3. In the **Name** field, enter the name you want to give your semantic type, UK Counties in this example.
4. In the **Description** field, enter List of counties in the United Kingdom.
5. In the **Type** drop-down list, select **Dictionary**.

You will indeed create this semantic type based on an exhaustive list of values.

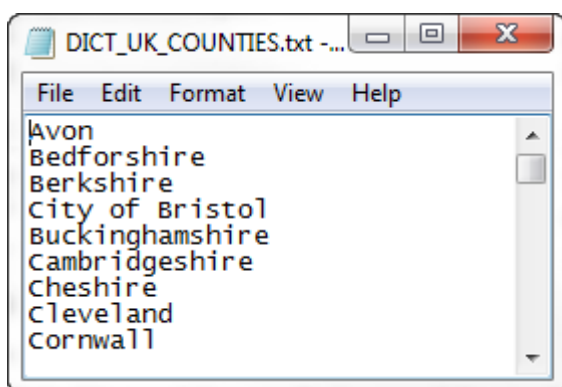
6. Keep the **Use for validation** switch activated.

Using a regular expression, a dictionary or a compound type for validation means that it will be used to define which values are considered right or wrong in a given column. The result of this validation process can be seen in the quality bar of each column in your datasets.

In any case, regular expressions or dictionary of values are used for data discovery, that calculates the matching percentage between the reference values and your data to define the semantic type of each column.

7. In the **Validation criterion** drop-down list, select the restriction rule that you want to apply, **Exact value** for example.
 - **Simplified text:** Punctuation, white spaces, case and accents are ignored during validation. For example, if Pâté-en-croûte is your reference value, pate-eN-cRoute will be considered valid but not Pâté n croûte.
 - **Ignore case and accents:** Case and accents are not taken into account during the validation. For example, if Pâté-en-croûte is your reference value, pate-en-croute will be considered valid but not pate en croute.
 - **Exact value:** The most restrictive validation rule. Data is considered as valid only if it is an exact match with the reference value.
8. To add the list of counties that will make up the UK Counties semantic type in the **Values** field, you can:
 - Manually add each value. Click the **plus** icon to enter a value, and click the **tick** icon to validate your change. Repeat for each county to add to the list.
 - Import file containing a plain text list of UK counties. Click the **import** button to select the file to upload. The file format is not important, as long as the content is plain text.

Retrieve the dict_uk_counties.txt file from the **Downloads** tab of the online version of this page, at <https://help.talend.com>.



Enter each different value on a separate line. Values that are on the same line and separated by a comma will be considered as synonyms.

When importing a list from a file, non-alphabetical values must be protected by quotes, otherwise the file will be rejected.

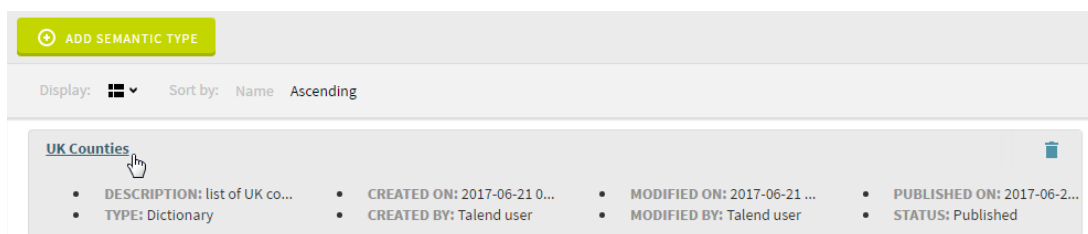
Duplication of values is not allowed. When manually adding values, a check is done. And when importing a file, a deduplication step is automatically performed.

The full list of counties has been added.

9. Click **Save and publish** to send the new semantic type to the Talend Dictionary Service server and make it available to the Talend Data Preparation users.

Clicking **Save as draft** means that the semantic type will be stored in Talend Dictionary Service, but will not be broadcasted to the Talend Web applications. This allows you to choose the moment when you want to make your semantic types public.

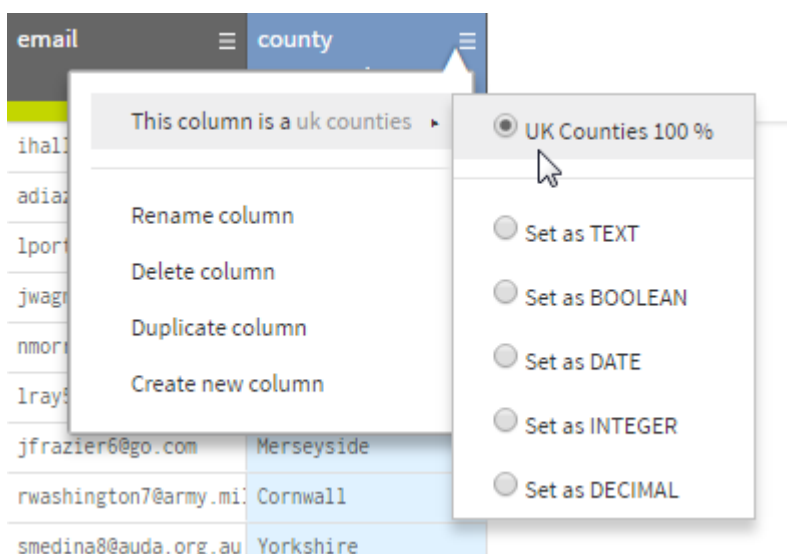
The **UK Counties** type is now available in the list of semantic types with the status set as **Published**.



The change in semantic types is instantly effective in Talend Data Preparation for every new dataset that you import. For existing datasets, you need to manually change the column type or reimport your dataset.

10. Go back to your dataset containing the counties names.
11. Click the menu icon in the **County** column header and select **this columns is a... > UK Counties**.

The column type now matches the newly created category.



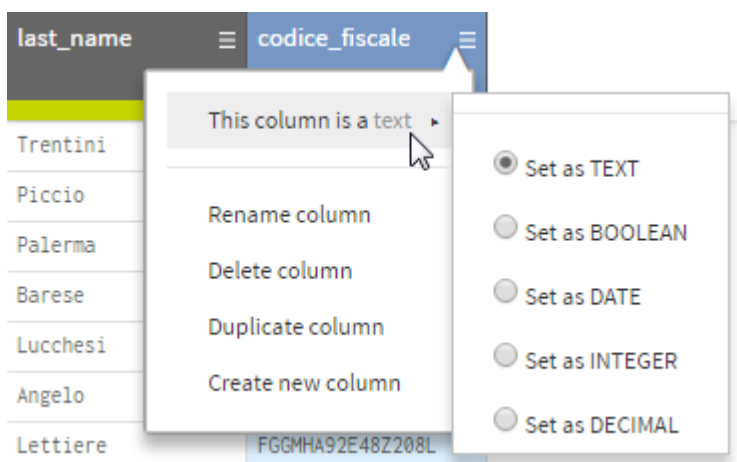
Your data is now matched with the **UK Counties** semantic type, that you manually created in Talend Dictionary Service. From now on, when importing new datasets containing names of British counties, they will automatically be matched with the proper type.

Adding a new regular expression-based semantic type through the UI

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 Preparation

In Talend Data Preparation, not every type of data can currently be matched with 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 clean some customer data, such as their names, email address, or their social security number. The semantic type for the column containing the social security number data will be set by default to `text`. This is not specific enough and you would like to create a new 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 Preparation so that your data can be matched with a proper type.

1. Open the **Semantic types** view from the left panel of the Talend Data Preparation homepage and click **Add semantic type**.
2. In the **Name** field, enter `codice fiscale`.
3. In the **Description** field, enter `Italian social security number`.
4. In the **Type** drop-down list, select **Regular expression**.
5. Keep the **Use for validation** switch activated.

Using a regular expression, a dictionary or a compound type for validation means that it will be used to define which values are considered right or wrong in a given column. The result of this validation process can be seen in the quality bar of each column in your datasets.

In any case, regular expressions or dictionary of values are used for data discovery, that calculates the matching percentage between the reference values and your data to define the semantic type of each column.

6. In the **Content** drop-down list, select the type of content that you want to validate, **Any character** in this case.

This option helps optimizing performances. Only the data that matches the selected type will be validated. You can choose to only validate **Alphabetic** or **Numeric** values against a regular expression, but because Italian social security numbers contain both, you have to select **Any character**.

7. In the **Validation pattern** field, enter `^[A-Z]{6}[0-9]{2}[A-Z][0-9]{2}[A-Z][0-9]{3}[A-Z]$`.

This regular expression is designed to match the Italian *codice fiscale*, which is an alphanumeric code of 16 characters. Data that matches that pattern in Talend Data Preparation will be identified as *codice fiscale*.

8. Click **Save and publish** to send the new semantic type to the Talend Dictionary Service server and make it available to the Talend Data Preparation users.

Clicking **Save as draft** means that the semantic type will be stored in Talend Dictionary Service, but will not be broadcasted to the Talend Web applications. This allows you to choose the moment when you want to make your semantic types public.

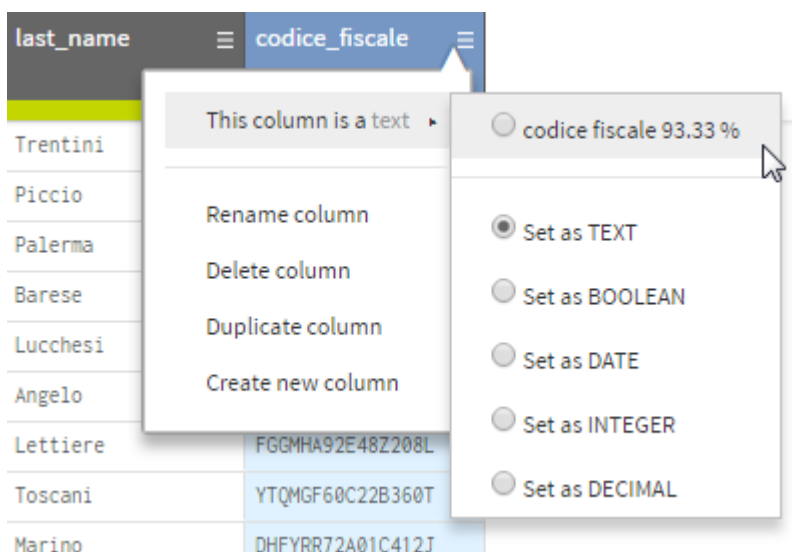
The **codice fiscale** type is now available in the list of semantic types with the status set as **Published**.

ADD SEMANTIC TYPE				
Display: Sort by: Name Ascending				
Civility				
• DESCRIPTION: Civility (multilingual)	• CREATED ON: 2016-12-08 15:38:49	• MODIFIED ON: N/A	• PUBLISHED ON: 2016-12-08 15:38:49	
• TYPE: Dictionary	• CREATED BY: Talend	• MODIFIED BY: N/A	• STATUS: Published	
codice fiscale				
• DESCRIPTION: Italian social security number	• CREATED ON: 2017-06-20 17:27:19	• MODIFIED ON: 2017-06-20 17:27:40	• PUBLISHED ON: 2017-06-20 17:27:40	
• TYPE: Regular expression	• CREATED BY: user@talend.com	• MODIFIED BY: user@talend.com	• STATUS: Published	

The change in semantic types is instantly effective in Talend Data Preparation for every new dataset that you import. For existing datasets, you need to manually change the column type or reimport your dataset.

9. Go back to your dataset containing the Italian social security numbers.
10. Click the menu icon in the **codice_fiscale** column header and select **this column is a...** > **codice fiscale**.

The column type now matches the newly created category.



Your data is now matched with the `codice_fiscale` semantic type, that you manually created in Talend Dictionary Service. From now on, when importing new datasets containing Italian social security numbers, they will automatically be matched with the proper type.

Adding a new compound semantic type

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

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.

In this example you need to prepare a file containing information about customers from the United States, the United Kingdom, Germany and France. One of the columns in this dataset contains postal codes from these different countries, and as a consequence, with different formats. In this case, Talend Data Preparation will apply the semantic type that matches the most with the values in the column, `US Postal code` for example. This will cause the rest of the data, German, French and British postal codes, to be considered invalid.

To make Talend Data Preparation more adapted to this situation, you will create a compound type, regrouping the several semantic types used to validate postal codes.

All the semantic types that you want to group under the compound type have been published.

1. Open the **Semantic types** view from the left panel of the Talend Data Preparation homepage and click **Add semantic type**.
2. In the **Name** field, enter `Postal code`.
3. In the **Description** field, enter `American, British, German and French postal codes`.
4. In the **Type** drop-down list, select **Compound type**.
5. Keep the **Use for validation** switch activated.

This compound type will be used to define which values are considered right or wrong when applied on a given column. The result of this validation process can be seen in the quality bar of each column in your datasets.

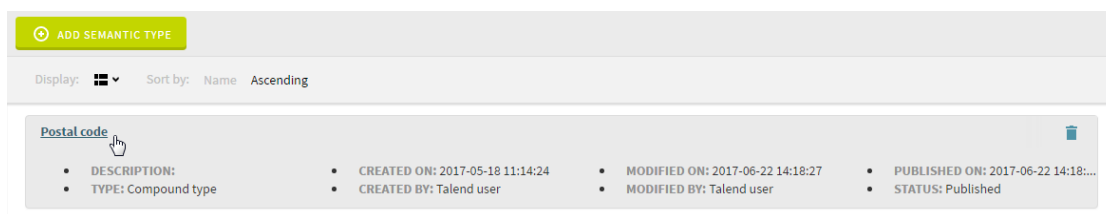
6. From the **Children types** drop-down list, select the semantic types you want to group under this `Postal code` compound type.



7. Click **Save and publish** to send the new compound type to the Talend Dictionary Service server and make it available to the Talend Data Preparation users.

Clicking **Save as draft** means that the semantic type will be stored in Talend Dictionary Service, but will not be broadcasted to the Talend Web applications. This allows you to choose the moment when you want to make your semantic types public.

The `Postal code` type is now available in the list of semantic types with the status set as **Published**.



The change in semantic types is instantly effective in Talend Data Preparation for every new dataset that you import. For existing datasets, you need to manually change the column type or reimport your dataset.

8. Go back to your dataset containing the postal codes from several countries.
9. Click the menu icon in the header of the column containing the postal codes and select **this columns is a... > Postal code**.

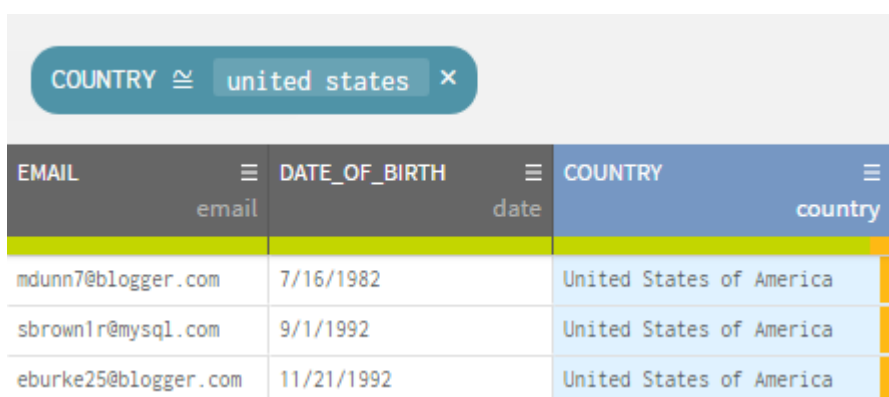
Your data is now matched with the `Postal code` compound type, that you manually created in Talend Dictionary Service. From now on, when importing new datasets containing postal codes, they will automatically be matched with the proper type.

Updating an existing semantic type through the UI

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

Predefined semantic types in Talend Data Preparation 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.



EMAIL	DATE_OF_BIRTH	COUNTRY
email	date	country
mdunn7@blogger.com	7/16/1982	United States of America
sbrown1r@mysql.com	9/1/1992	United States of America
eburke25@blogger.com	11/21/1992	United States of America

The problem here is that **United States of America** is not one of the expected value for the `country` semantic type in Talend Data Preparation. 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 Preparation.

1. Open the **Semantic types** view from the left panel of the Talend Data Preparation homepage.
2. From the list of existing semantic types, click the **Country** type to open it.
In this window, all the parameters of the semantic type can be modified, including the list of entries used to discover or validate data.
3. In the **Values** list, point your mouse over the **United States** entry and click the pen icon that is displayed on the right.
4. Right after **United States**, enter `United States of America` as second value, separated by a comma.
5. Click the tick icon to validate your change.

Those two values, that were entered in the same row, are now set as synonyms. As a consequence, **United States of America** will now be considered a valid value for the `country` semantic type.

6. Click **Save and publish** to propagate the change in Talend Dictionary Service and make it available to the Talend Data Preparation users.

The change in semantic types is instantly effective in Talend Data Preparation for every new dataset that you import. For existing datasets, you need to duplicate the column or reimport your dataset.

7. Go back to your dataset with the column containing the customers countries.
8. Duplicate the column with the updated semantic type applied, **Country** in this case.

You can see in the quality bar under the column header that there is no invalid values anymore.

COUNTRY	COUNTRY_COPY
country	country
Central African Republic	Central African Republic
Czech Republic	Czech Republic
United States of America	United States of America
Nicaragua	Nicaragua

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.

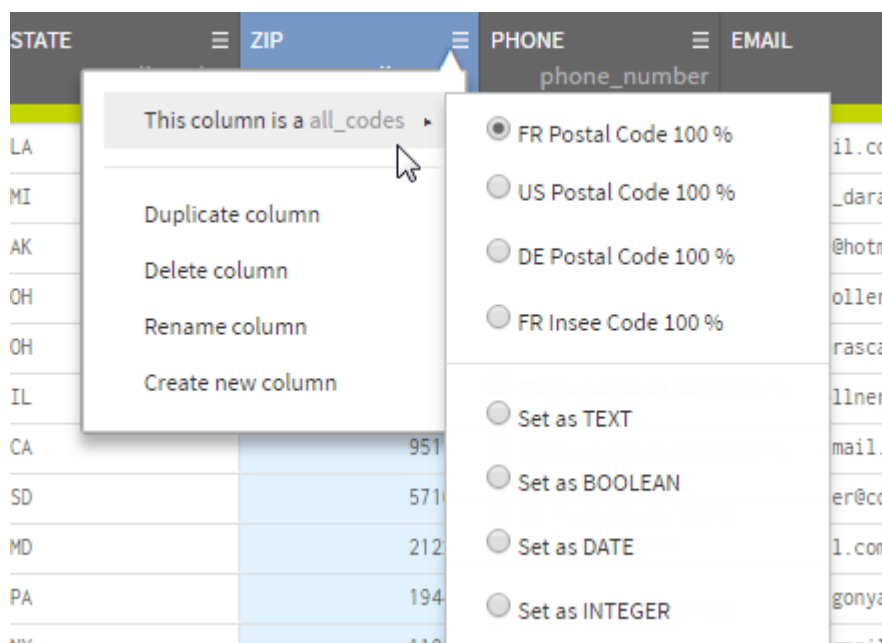
Removing a semantic type through the UI

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

The variety of semantic types that are present by default in Talend Data Preparation may not apply to your business context. 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.

In this example, the **ZIP** column of the dataset can be matched with at least four 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 Preparation, and from now on, ZIP codes will only be validated against the `US Postal Code` semantic type.

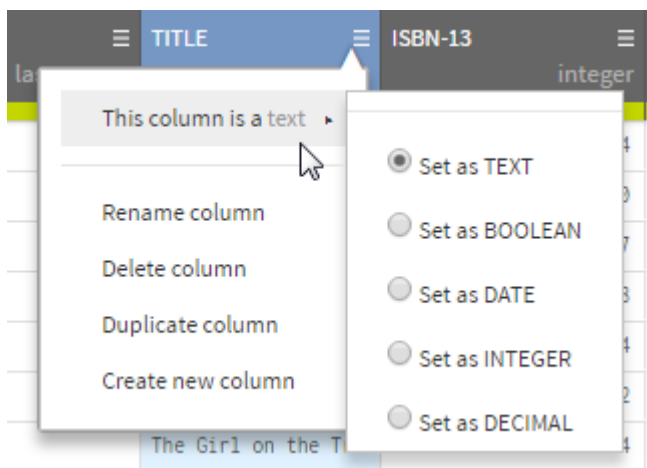
1. From the left panel of the Talend Data Preparation homepage, open the **Semantic Types** view.
2. In the list of existing semantic types, look for **FR Postal Code**.
3. To delete it, point your mouse over the semantic type and click the garbage bin icon that is displayed on the right.
4. Repeat the last two steps to delete the **FR Insee Code** and the **DE Postal Code**.

You have deleted the other semantic types compatibles with five-digit numbers. From now on, when adding new datasets, only **US Postal Code** will be proposed as semantic type for the columns containing Zip codes.

If you remove a semantic type that is used in one or more datasets, the relevant columns will switch back to the **text** category.

Enriching the semantic types libraries

When you add a dataset, Talend Data Preparation automatically suggests one of the supported semantic types for each column. If the semantic type proposed by Talend Data Preparation for one column is not the desired one, you can manually change it by clicking the white arrow in the column header.



This allows you to choose among the list of semantic types present in Talend Data Preparation by default. See [Predefined Semantic Types](#) on page 146 for more information. You can go further by creating your own semantic types, as well as updating or deleting the existing ones, so that Talend Data Preparation speaks your business language.

The semantic types modifications are made using Talend Dictionary Service. This tool stores all the semantic libraries used in various Talend products, including Talend Data Preparation. All the changes that you make in the Talend Dictionary Service server will be instantly available in Talend Data Preparation. The availability of Talend Dictionary Service depends on the license you have.

In Talend Dictionary Service, the semantic types are divided into three main categories:

- The **DICT** type, based on an open or closed list of values.
- The **REGEX** type that compares your data against a preselected regular expression.
- The **COMPOUND** type, under which you can group several existing types

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 Preparation, you must fulfill the following prerequisites:

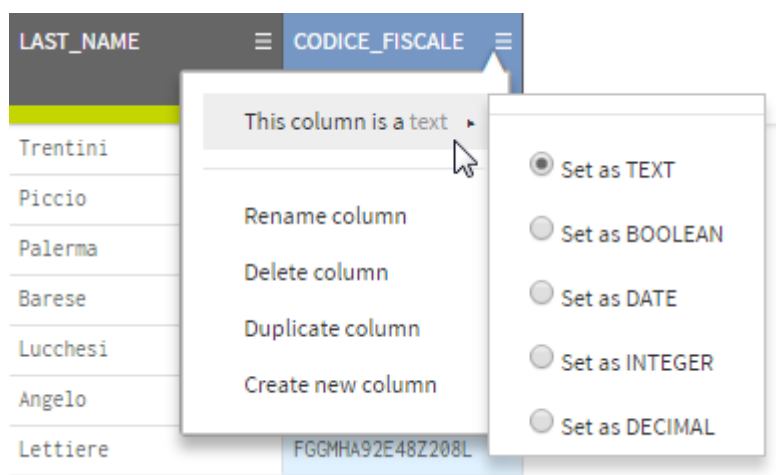
- Talend Dictionary Service is installed and running.
- Talend Administration Center is installed and running.
- Your Talend Administration Center user type is either **Master Data Management** or **Data Quality**
- The **Data Preparation User** check box is selected for your user in Talend Administration Center with any of the three possible roles set in the **Data Preparation Role** field.
- In the `<install_folder>\dataprep\config\application.properties` file, the `dataquality.semantic.update.enable` property is set as `true`.

Adding a new regular expression-based semantic type through command line interface

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 Preparation.

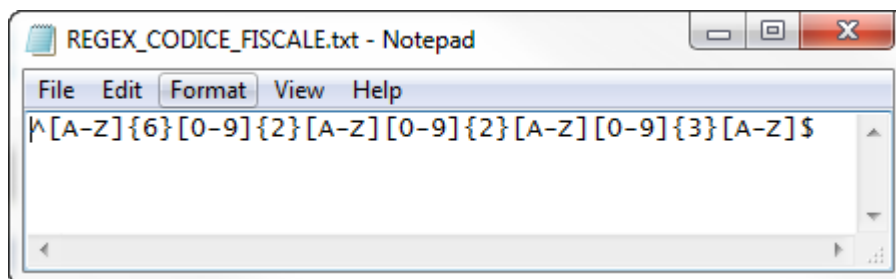
In Talend Data Preparation, not every type of data can currently be matched with one of the predefined semantic types. Italian social security numbers, also known as *codice fiscale*, are currently not recognized for example.

Let's say that you work for an Italian company, only dealing with Italian customers. In this example, you need to clean some customer data, such as their names, email address, or their social security number. The semantic type for the column containing the social security number data will be set by default to `text`. 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 Preparation so that your data can be matched with 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. Data that matches that pattern in Talend Data Preparation will be identified as codice fiscale.

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.
4. Using the cd command, go to the <Dictionary_Service_Path>/command-line folder.
5. To create the new codice_fiscale semantic type in Talend Dictionary Service and configure its different parameters, execute the following command 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.

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.

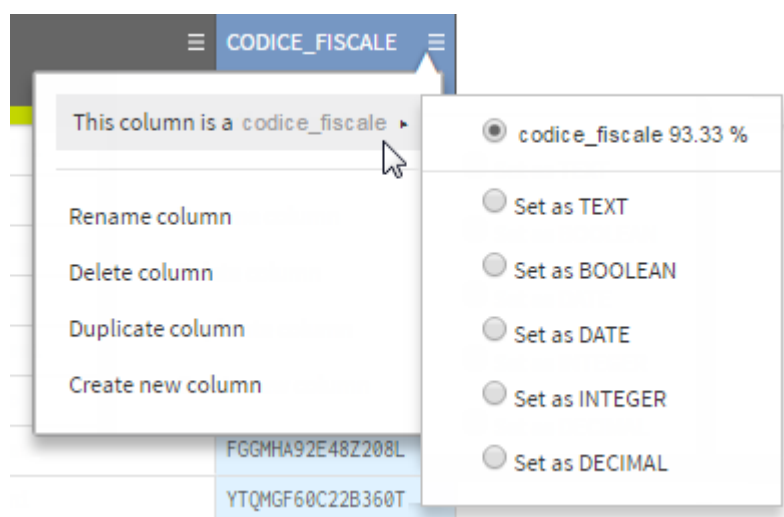
The codice_fiscale semantic type is now added to the list of categories in Talend Dictionary Service.

6. Go back to Talend Data Preparation and open your dataset with the column containing the social security numbers.

The change in semantic types is instantly effective in Talend Data Preparation for every new dataset that you import. For existing datasets, you need to manually change the column type.

7. To apply the new codice_fiscale semantic type to your column, click the white arrow next to the column name.
8. Click **This column is a... > codice_fiscale**

The column type now matches the newly created category.



Your data is now matched with the `codice_fiscale` semantic type, that you manually created in Talend Dictionary Service. From now on, when importing new datasets containing Italian social security numbers, they will automatically be matched with the proper 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:

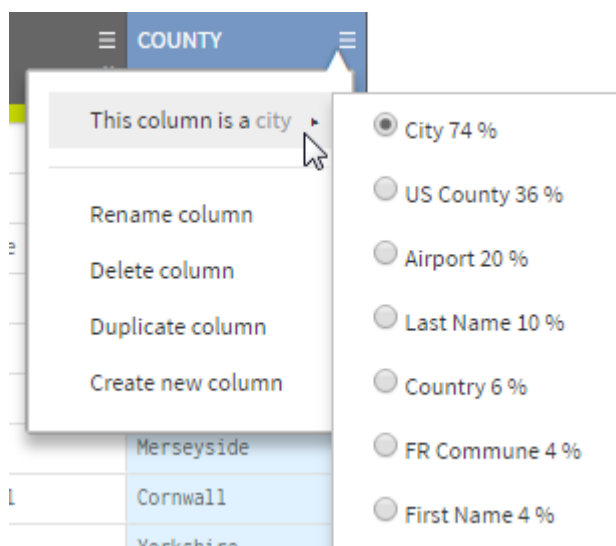
- `category_manager.bat -h` command for Windows.
- `./category_manager.sh -h` for Linux.

Adding a new dictionary-based semantic type through command line interface

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

In Talend Data Preparation, 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 clean some customer data, such as their names, email address, or the county they live in. The semantic type for the column containing the counties data will be set by default to `city`. Some of the data may actually match names of cities, but you want to add a semantic type that is more 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 Preparation so that your data can be matched with a proper type.

1. Create a .txt file containing the exhaustive list of the British counties and save it as `DICT_UK_COUNTIES.txt`.

You must enter only one entry per line.



Unlike an open dictionary which purpose is to identify data, this exhaustive list will act as a closed dictionary of values to identify and validate data in Talend Data Preparation. Data that exactly matches one of the listed values will be 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
4. Using the `cd` command, go to the `<Dictionary_Service_Path>/command-line` folder.
5. To create the new `UK_counties` semantic type in Talend Dictionary Service and configure its different parameters, execute the following command 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\DICTIONARIES\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/DICTIONARIES\UK_COUNTIES.txt` 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.

The `-cmpl` attribute stands for completeness, and is used to determine if the dictionary you are adding is an open or a closed dictionary. It is set to `false` by default but in this case, it must be set to `true`.

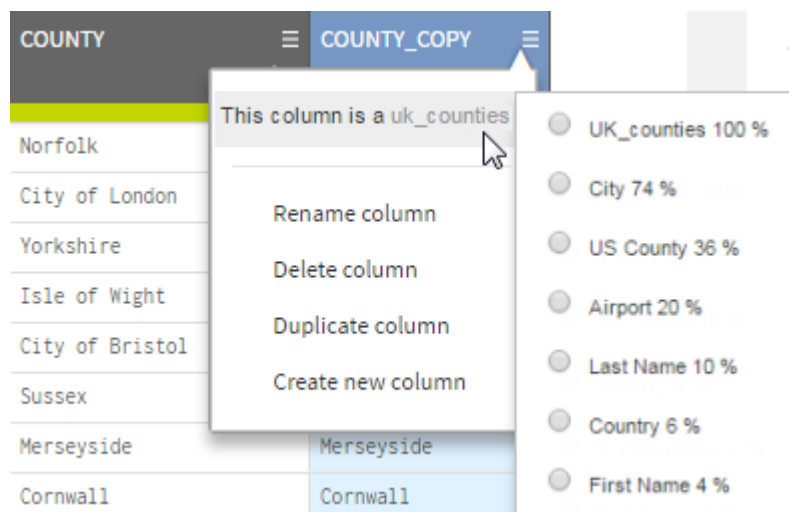
The `UK_counties` semantic type is now added to the list of categories in Talend Dictionary Service.

6. Go back to Talend Data Preparation and open the dataset with the column containing the counties names.

The change in semantic types is instantly available in Talend Data Preparation, but you need to manually refresh the column to make it visible in your existing datasets and preparations.

7. To make the changes in semantic types active, you can either:
 - import your dataset again.
 - make a copy of the column which semantic type you want to update, **COUNTY** in this example.

The column type now matches the newly created category.



Your data is now matched with the `UK_counties` semantic type, that you manually created in Talend Dictionary Service. From now on, when importing new datasets containing names of British counties, they will automatically be matched with the proper 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:

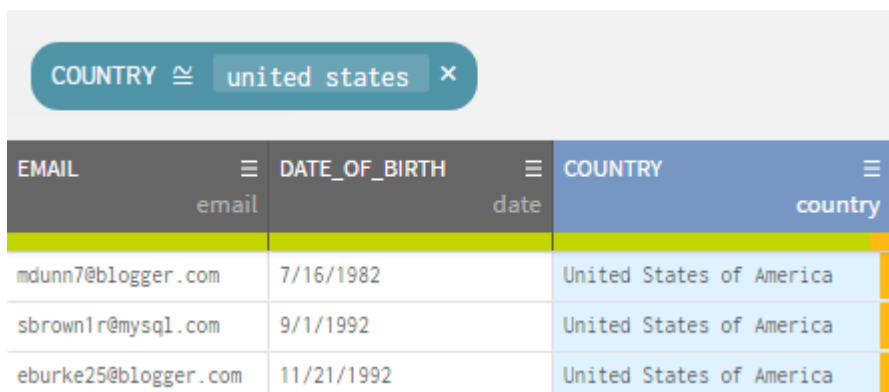
- `category_manager.bat -h` command for Windows.
- `./category_manager.sh -h` for Linux.

Updating an existing semantic type through command line interface

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

Predefined semantic types in Talend Data Preparation 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.



EMAIL	DATE_OF_BIRTH	COUNTRY
email	date	country
mdunn7@blogger.com	7/16/1982	United States of America
sbrown1r@mysql.com	9/1/1992	United States of America
eburke25@blogger.com	11/21/1992	United States of America

The problem here is that **United States of America** is not one of the expected value for the `country` semantic type in Talend Data Preparation. 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 Preparation.

1. Open a command prompt window
2. Using the `cd` command, go to the `<Dictionary_Service_Path>/command-line` folder.
3. 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.

4. 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.



5. Go back to Talend Data Preparation and open your dataset with the column containing the customers countries.

The change in semantic types is instantly available in Talend Data Preparation, but you need to manually refresh the column to make it visible in your existing datasets and preparations.

6. To make the change in the countries list active, you can either:

- import your dataset again.
- make a copy of the column which semantic type you want to update, **COUNTRY** in this example.

You can see in the quality bar under the column header that there is no invalid values anymore.

COUNTRY 	COUNTRY_COPY 
country	country
Central African Republic	Central African Republic
Czech Republic	Czech Republic
United States of America	United States of America
Nicaragua	Nicaragua

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 through command line interface

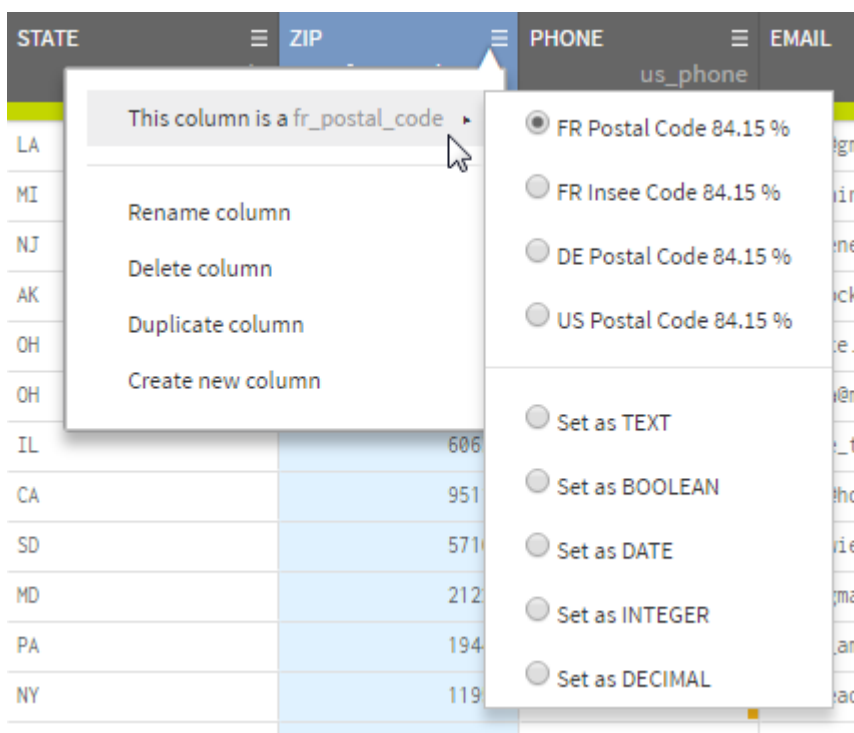
You can delete a semantic type in Talend Dictionary Service to remove it from the list of recognized data types in Talend Data Preparation.

This applies to both predefined semantic types, as well as custom semantic types.

The variety of semantic types that are present by default in Talend Data Preparation may not apply to your business context. 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.

Talend Data Preparation tends to automatically match five-digit number with French ZIP codes. 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. Always having the wrong semantic type in your columns containing ZIP codes can quickly become annoying.

In this example, the **ZIP** column of the dataset you are preparing can be matched with at least four 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 Preparation, and five-digit numbers will automatically be identified as US ZIP codes from now on.

1. Open a command prompt window.
2. Using 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 five-digit numbers will not be associated with French ZIP codes anymore.

5. Repeat this operation to remove the other semantic types that match five-digit numbers:

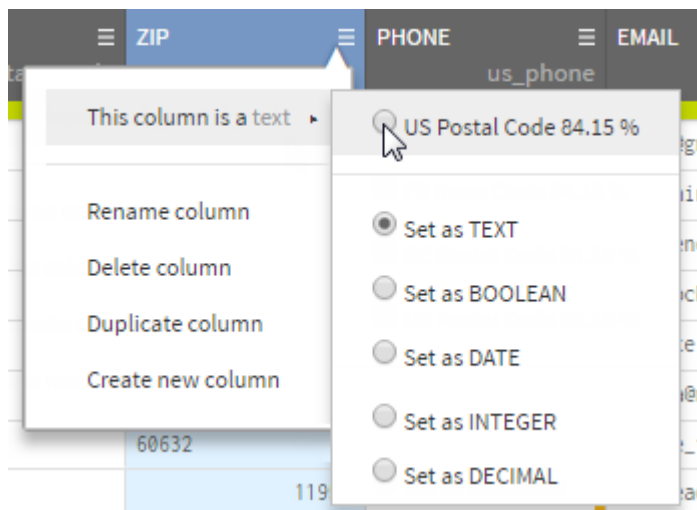
- `DE_POSTAL_CODE`
- `FR_INSEE_CODE`

6. Go back to your preparation with the column containing ZIP codes in Talend Data Preparation.

The change in semantic types is instantly available. Because you deleted the semantic type that was used until now, the **ZIP** column is automatically defined as `text`.

7. To set the proper semantic type to the column, click the white arrow in the column header.

8. Point your mouse over **This column is a text** and select **US Postal Code**.



This time, the data from the **Zip** can only be matched with the `US_POSTAL_CODE` semantic type. You have deleted all the semantic types compatibles with five-digit numbers but one. From now on, when adding new datasets, this type of data will be identified as US postal codes.

To display a list of all the available commands in Talend Dictionary Service, enter the `category_manager.bat -h` command for Windows or `./category_manager.sh -h` for Linux

Filtering values manually

In order to have a more specific idea of the data contained in your dataset or in order to perform functions on a certain subset of data, you can create a filter on your data.

This example uses a dataset with typical customer information, such as their names, age, email or state they live in. You are going to manually enter some filters to only display the male customers from California.

	ID	FIRST_NAME	LAST_NAME	EMAIL	GENDER	STATE
	integer	first_name	last_name	email	gender	us_state
1	1	Christina	Fox	cfox0@istockphoto.com	Female	New York
2	2	Aaron	Stewart	astewart1@sun.com	Male	Pennsylvania
3	3	Henry	Butler	hbutler2@yellowbook.com	Male	Florida
4	4	Jeremy	Morris	jmorris3@irs.gov	Male	Pennsylvania
5	5	Douglas	Elliott	delliott4@vk.com		West Virginia

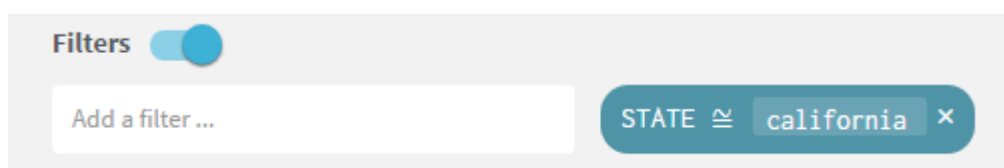
1. In the **Add a filter** field on the top left of the grid, start typing the value you want for your filter, `california` in this example. Talend Data Preparation suggests columns containing this value.



The suggestions are based on the data that is contained in the sample.

2. Select **california in state** to only display the entries corresponding to this location.

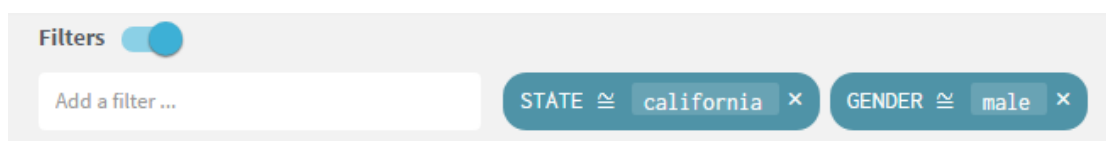
You can see in the Filter bar that the filter has been correctly applied.



Filter badges can be edited to search for any value.

You will now apply another filter, to isolate the male customers, among the ones already filtered.

3. In the **Add a filter** field, start typing male.
4. Select **male in gender** to add this filter to the previous one.



The grid now only displays the data corresponding to those two filters

5. In the functions panel, click a function to execute it on the data you filtered, **Keep these Filtered Rows** for example.
6. In the filter bar, click the cross in each individual filter or click the garbage bin icon to clear the filters and display the whole dataset again.

You have filtered your data to isolate a specific customer group and you can start applying function and work on this sample only.

You can apply filters manually, or use the charts panel to create even more complex filters.

Filtering values using the quality bar

The quickest way to identify and filter incorrect data is to use the quality bar.

ZIP	PHONE	EMAIL
fr_postal_code	us_phone	email
70116	504-621-8927	jbutt@gmail.com
48116	810-292-9388	josephine_darakjy@da
8014	856-636-	art@venere
99501	907-385-4412	lpaprocki@hotmail.co

Under each column is a quality bar that displays the amount of fields that have correct data, incorrect data or empty fields. Each category is represented by a color:

- Green for data that matches the cell format.
- White for empty cells.
- Orange for data that does not match the cell format.

From this quality bar, you can either choose to directly remove all the rows with empty or invalid data from column, or just select them and apply a filter on your data

Filters ☒

Add a filter ...

PHONE : rows with invalid values ✕

EMAIL : rows with empty values ✕

	fr_postal_code	PHONE	EMAIL	SUBDATE
	fr_postal_code	us_phone	email	date
304	92234	760		24-Jul-2012
441	77301	936-751		27-Oct-2013
483	11530	516-393-9		15-Nov-2010

Removing empty and invalid rows

Using the quality bar is a quick way to remove the rows containing invalid or empty records for a given column.

Let's take the example of a dataset containing some customer data. One of the column contains email addresses but some of the entries are either invalid or empty.

EMAIL	SUB
email	date
jbutt@gmail.com	17-11
josephine_darakjy@da	15-11
art@venere	28/1
lpaprocki@hotmail.co	24-11
donette.foller@cox.ne	17-11
simona@morasca.com	13-11

You are going to use the quality bar to directly delete all the rows containing empty or invalid values for this column

1. Click the white part of the quality bar, under the column header.
2. In the drop-down menu, click **Delete the rows with empty cells**.

The empty cells of the column have been deleted and only the invalid values, represented by the orange bar, remain.



3. Click the orange part of the quality bar.
4. In the drop-down menu, click **Delete the rows with invalid cells**.

Your column is now cleaned of all invalid data or empty cells.



Filtering values using charts

The **Chart** tab shows a graphical representation of your data. It is also a quick and easy way to apply filter on your data.

According to the type of data that you select, the type of graphical representation in the tab will be different:

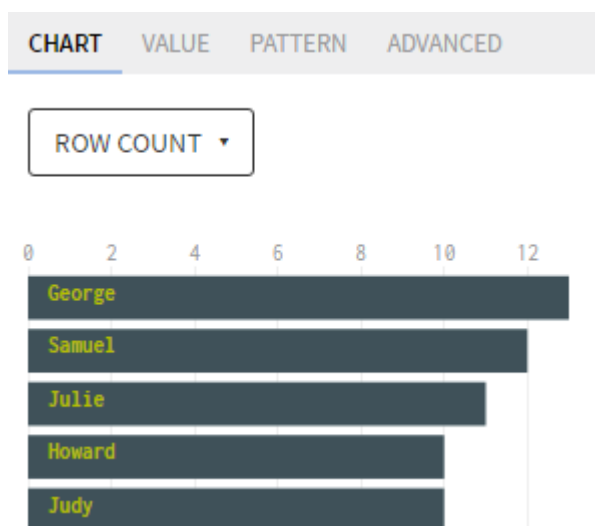
- Vertical bar charts for numerical data
- Horizontal bar charts for text data
- US map for the two-letter US State codes
- World map for the two-letter country codes

This example uses a dataset with typical customer information, such as their names, gender, email or the country they live in.

	ID	FIRST_NAME	LAST_NAME	GENDER	EMAIL	ISO2_COUNTRY_CODE
	integer	first_name	last_name	gender	email	country_code_iso2
1	1	Emily	Graham	Female	egraham0@princeton.edu	TH
2	2	Ashley	Little	Female	alittle1@bloglovin.com	PH
3	3	Nicholas	Peters	Male	npeters2@umn.edu	CN
4	4	Andrew	Romero	Male	aromero3@blog.com	BA
5	5	Samuel	Williams	Male	swilliams4@accuweather.com	EG

1. Select a column containing text data you want to filter, **FIRST_NAME** for instance.

The horizontal bar chart showing the most common occurrences of first names is displayed in the chart tab.



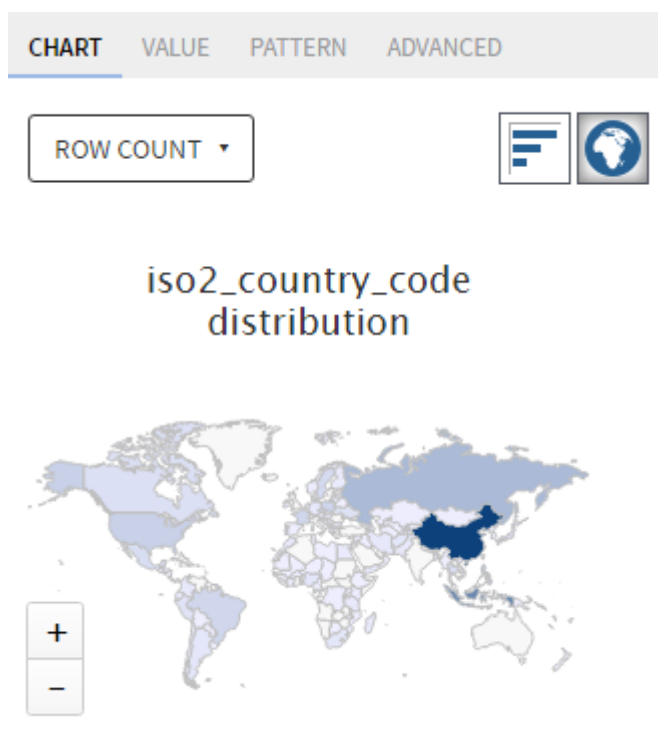
- Click the top bar to apply a filter on the most common first name.

The preparation now only displays the rows with this first name.

You can also use **Ctrl + Click** or **Shift + Click** to select multiple values at the same time and apply a more complex filter.

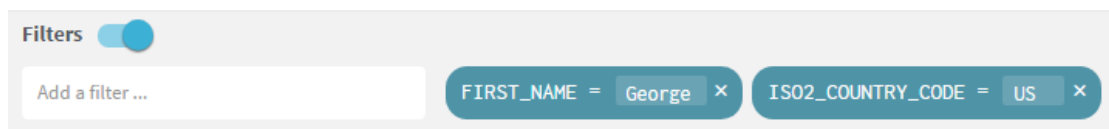
- Select the **ISO2_COUNTRY_CODE** column.

This time, the data is displayed in the form of a world map. The more occurrences of a country there is, the darkest this country will be on the map.



You can alternate between the world map view and the usual bar chart view by clicking the icons on the top right of the chart tab.

- Click the United States directly on the map to add this filter to the previous one.



The grid now only displays the data corresponding to those two filters.

5. In the **Functions panel**, click a function to execute it on the data you filtered, **Delete these Filtered Rows** for example.
6. In the filter bar, click the cross in each individual filter or click the garbage bin icon to clear the filters and display the whole dataset again.

Finding similar values

If you want to find and filter some text that looks alike, in order to fix typos for example, you can use the **Match Similar Text** function.

This function creates a new column with the value **true** if the pattern matches and **false** if it does not.

1. Select the text column where you want to find similar text.
2. In the **Functions panel**, type `Match Similar Text` and click the result to open the options for the associated function.
3. Fill in the options according to your needs.

The **Reference** field corresponds to some text you enter, and the **Fuzziness** field corresponds to the number of characters that can be added, removed or different from the **Reference**. This number is called the Levenshtein distance.

Note that the **Reference** field is case sensitive. In this example, the **Reference** text is `new` and the Levenshtein distance (**Fuzziness**) is 1.

In this example, the function would match words such as "few", "now", "net" or "news", but not "bow", "nap" or "led".

4. Click the **Submit** button to apply the function with the selected options.

This creates a new column with the value **true** if the pattern matches and **false** if it does not.

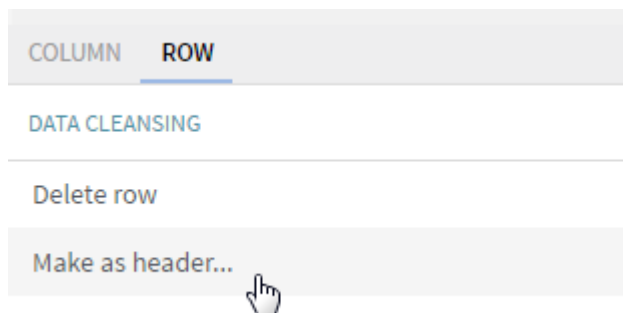
For more information on the Levenshtein distance, see https://en.wikipedia.org/wiki/Levenshtein_distance

Setting a line as header

Talend Data Preparation automatically sets the first row as a header for your dataset.

If this first row is not the header in your dataset, you can use the **Make as header** function to specify the header.

1. Click the line you want to set as the header.
2. In the **Functions panel**, click the **Make as header** function.





The selected row is now the header of the dataset.

Swapping column content

In the case where the content of two columns has been inverted, you can use the **Swap columns** function.

This function allows you to swap the data contained in two columns without having to rename the columns and modify their type.

1. Click one of the columns with the incorrect data.

DIVISION 	TEAM 
text	airport
Boston Bruins	Atlantic
Buffalo Sabres	Atlantic
Detroit Red Wings	Atlantic
Florida Panthers	Atlantic

2. In the **Functions panel**, type `Swap columns`.
3. In the drop down list, select the name of the column with which you want to swap the data.

Swap columns...

Column:
Team ▼

SUBMIT

4. Click **Submit**.

The content of the two columns have been swapped.

DIVISION	TEAM
airport	text
Atlantic	Boston Bruins
Atlantic	Buffalo Sabres
Atlantic	Detroit Red Wings
Atlantic	Florida Panthers

Working with the data

Applying functions on multiple columns

Rather than applying the same function to different columns one after the other, you can perform actions on several columns at the same time.

Let's take the example of a dataset containing several columns with date data, each of them set in a different format.

	SUBSCRIPTION DATE	LAST RENTAL	BIRTHDAY
	date	date	date
1	31/08/2010	03/08/2017	1981-12-12
2	25/10/2015	12/09/2016	1964-04-16
3	21/12/2016	06/08/2016	1977-12-08
4	25/12/2013	08/11/2016	1989-07-08
5	28/07/2016	06/10/2016	1984-09-14

The different date formats used are the following:

- dd/mm/yyyy for the **Subscription date** column
- mm/dd/yyyy for the **Last rental** column
- yyyy-mm-dd for the **Birthday** column

You are going to harmonize the date format, and set these columns to the French standard: dd/mm/yy.

1. Click the **Subscription date** column.
2. To select the two remaining columns, you have two options:
 - While pressing the **Ctrl** key, click the **Last rental** column and the **Birthday** column.
 - While pressing the **Shift** key, select the **Birthday** column.

The **Shift + click** option allows you to select all the columns between your first selection and your last.

When selecting multiple columns, no charts are available for the data.

3. In the **Functions panel**, type `Change Date Format`.
4. From the **New format** drop-down list, select **French standard**.
5. Click **Submit** to apply the function on the three columns.

The date format for the three selected columns is now set to the French standard.

In addition, three new steps are added to the recipe, one for each column.

	≡ SUBSCRIPTION DATE ≡ date	≡ LAST RENTAL ≡ date	≡ BIRTHDAY ≡ date
1	31/08/10	08/03/17	12/12/81
2	25/10/15	09/12/16	16/04/64
3	21/12/16	08/06/16	08/12/77
4	25/12/13	11/08/16	08/07/89
5	28/07/16	10/06/16	14/09/84

Changing the date format

As the date formats used across the world are not the same, you may need to change the format used in a column containing dates.

1. Select a column containing dates.
2. In the **Functions panel**, click **Change Date Format** in the **Suggestion** part to open the options for the associated function.
3. In the **New format** list, select the date format you want to apply.

Change date format...

Current format:
I don't know, best guess ▼

New format:
custom ▼

Your format:
dd/MM/yyyy

SUBMIT

[Learn more ...](#)

See [Date Formats](#) on page 120 for more information regarding the available options.

4. In the **Your Format** field, type `dd/MM/yyyy` to change the date from an American format to a French one.

For example, this will change 12/25/2015 to 25/12/2015.

5. Click **Submit**.

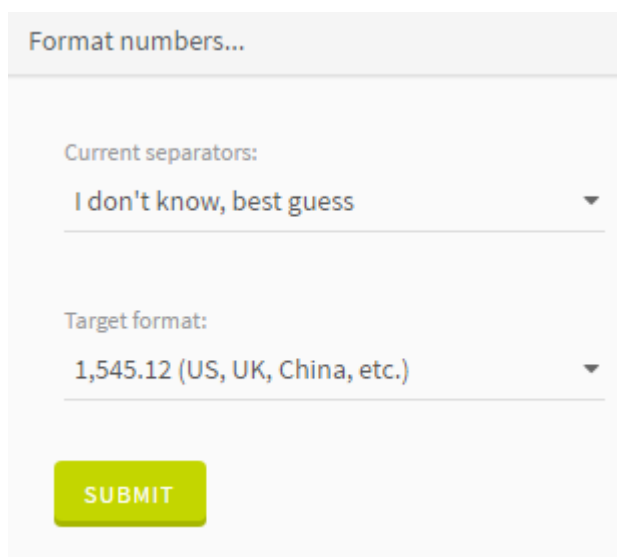
The date format is changed in the selected column.

Changing the number format

As the number formats used across the world are not the same, you may need to change the format used in a column containing numbers.

By default, the number format used in Talend Data Preparation corresponds to the format used on your computer.

1. Select the column for which you want to change the number format.
2. In the **Functions panel**, type `Change Number Format` and click the result to open the options for the associated function.
3. In the **Current separators** list, select the number format used in the selected column. You can:
 - use a predefined separator
 - define it yourself with **custom**
 - let Talend Data Preparation automatically guess the format



The image shows a 'Format numbers...' dialog box. It has two dropdown menus. The first, labeled 'Current separators:', has the text 'I don't know, best guess' and a downward arrow. The second, labeled 'Target format:', has the text '1,545.12 (US, UK, China, etc.)' and a downward arrow. At the bottom is a yellow button labeled 'SUBMIT'.

4. In the **Target Format** list, choose the number format you want to use in the selected column.
5. Click **Submit**.

The number format is changed in the selected column.

Changing the case to lower case

Sometimes, you may have to change the case of some text to lower case. This can be useful if you want to append this text to some other text.

1. Select a column containing text you want to change to lower case.
2. In the **Functions panel**, type `Change to lower case` and click the result to execute the associated function.
3. Repeat this action for every column you want to modify.

The text contained in the selected column is all in lower case.

Clearing cells corresponding to a pattern

You can remove the content of cells that match a value or a pattern that you have defined.

1. Select the column in which you want to remove the content that matches your pattern.
2. In the **Functions panel**, type `Clear on matching value` and click the result to open the options for the associated function.
3. Enter the value that you want to match and the operator you want to use.

DATA CLEANSING

Clear on matching value...

Value:

IR

SUBMIT

[Learn more ...](#)

4. Click the **submit** button to apply the function with the selected options.

The content of the cells that correspond to the pattern that you have defined is removed.

Comparing dates

Comparing dates allows you to know, for example, if the dates contained in the selected column are before or after a given date.

1. Select a column containing dates.
2. In the **Functions panel**, type `Compare Dates` and click the result to open the options for the associated function.
3. In the **Compare mode** list, choose the comparison mode. You can compare the column with another column or with a given date.

Compare dates...

Compare mode:

equals

Use with:

Constant

Constant:

2017-04-27 17:29

SUBMIT

The dates contained in the selected column are compared, either with a date you entered or with dates contained in another column. The result of the comparison is written in a new column, **true** when the condition is matched, **false** when the condition is not.

Changing the calendar

Talend Data Preparation allows you to easily switch from one type of calendar to the other with the **Convert dates** function.

In this example, you are preparing some customer data that needs to be sent to one of your Japanese clients. Your dataset includes client information such as names, phone numbers, email addresses, but also the date on which they subscribed to a specific service.

	FIRST_NAME	LAST_NAME	AGE	SUBDATE	EMAIL
	text	text	phone_number	date	email
4	Lenna	Paprocki	45-49	24/11/2013	lpaprocki@hotmail.com
5	Donette	Foller	25-34	17/04/2012	donette.foller@cox.net
6	Simona	Morasca	50-55	13/04/2016	simona@morasca.com
7	Mitsue	Tollner	35-44	07/06/2009	mitsue_tollner@yahoo.com
8	Leota	dilliard	25-34	04/12/2008	leota@hotmail.com
9	Sage	Wieser	25-34	20/04/2013	sage_wieser@cox.net

You can see that in this dataset, the subscription dates have been entered using the dd/MM/yyyy format of the Gregorian calendar. Because the preparation is aimed at the Japanese market and you want to make it more adapted to this target, you will convert these dates to the Japanese calendar.

1. Click the header of the **Subscription date** column to select its content.
2. In the **Functions panel** select **Convert Dates**.

The menu for the corresponding function opens.

3. In the **Source calendar** drop-down list, select **Gregorian calendar**.
4. In the **Target calendar** drop-down list, select **Japanese calendar**.

Convert date...

Current calendar type:
Gregorian calendar ▼

New calendar type:
Japanese calendar ▼

SUBMIT

5. Click **Submit** to apply the function on the column.

The subscription dates are now using the Japanese calendar, while keeping the dd/MM/yyyy format, and you preparation is more adapted to your client.

	FIRST_NAME	LAST_NAME	AGE	SUBDATE	EMAIL
	text	text	phone_number	date	email
4	Lenna	Paprocki	45-49	24/11/0025	lpaprocki@hotmail.com
5	Donette	Foller	25-34	17/04/0024	donette.foller@cox.net
6	Simona	Morasca	50-55	13/04/0028	simona@morasca.com
7	Mitsue	Tollner	35-44	07/06/0021	mitsue_tollner@yahoo.com
8	Leota	dilliard	25-34	04/12/0020	leota@hotmail.com
9	Sage	Wieser	25-34	20/04/0025	sage_wieser@cox.net

Converting a calendar to Julian Day

The **Convert dates** function also allows you to switch to a calendar based on a day count, such as Julian Day.

Let's take the example of a dataset containing astronomical observations about meteor events and the date they were recorded. The date records are using the ISO 8601 format of the Gregorian calendar. You will use the appropriate function to change the calendar used for your date data from the standard ISO 8601 format, to Julian Day. The Julian Day calendar is commonly used in the astronomy field, and would be more fitting when outputting this data.

	Date	Time	Latitude	Longitude
	date	text	text	text
1	2015-04-30	10:21:01 AM	48.7S	139.1E
2	2015-04-21	01:42:51 AM	37.7N	39.6W
3	2015-04-08	04:06:31 AM	25.5S	51.5E
4	2015-04-03	01:39:38 AM	8.4N	157.9W
5	2015-03-30	09:33:52 PM	36.1S	5.5W
6	2015-03-18	12:04:50 AM	5.4S	159.3E

1. Click the header of the **Date** column to select its content.
2. In the **Functions panel** select **Convert Dates**.
The menu for the corresponding function opens.
3. In the **Source calendar** drop-down list, select **Gregorian calendar**.
4. In the **Target calendar** drop-down list, select **Julian Day**.

Convert date...

Current calender type:
Gregorian calendar ▼

New calender type:
Julian day ▼

SUBMIT

5. Click **Submit** to apply the function on the column.

The values contained in the **Date** column have been converted to Julian Days and now display the number of days that have passed in this specific calendar.

	Date	Time	Latitude	Longitude
	date	text	text	text
1	2457142	10:21:01 AM	48.7S	139.1E
2	2457133	01:42:51 AM	37.7N	39.6W
3	2457120	04:06:31 AM	25.5S	51.5E
4	2457115	01:39:38 AM	8.4N	157.9W
5	2457111	09:33:52 PM	36.1S	5.5W
6	2457099	12:04:50 AM	5.4S	159.3E

Calendar types

Here are the different calendar types that can be used on your data in Talend Data Preparation.

Calendars used in different cultures around the world.

Calendar	Description
Gregorian	https://en.wikipedia.org/wiki/Gregorian_calendar
Hijri	https://en.wikipedia.org/wiki/Islamic_calendar
Japanese	https://en.wikipedia.org/wiki/Japanese_calendar
Minguo	https://en.wikipedia.org/wiki/Minguo_calendar
ThaiBuddhist	https://en.wikipedia.org/wiki/Buddhist_calendar

Calendars that are based on a day count, starting from a specific reference date.

Calendar	Description
Julian day	https://en.wikipedia.org/wiki/Julian_day
Modified Julian day	https://en.wikipedia.org/wiki/Julian_day
Epoch day	https://en.wikipedia.org/wiki/Unix_time
Rata die	https://en.wikipedia.org/wiki/Rata_Die

Deleting lines containing a specific value

You can delete lines that match a value that you have defined.

1. Create a filter for your values. See [Filtering values manually](#) on page 54 for more details.
2. In the **Functions** panel, select **Delete these filtered rows**.

SUGGESTIONS

Delete these filtered rows

The lines corresponding to the filter you have defined are deleted.

Deleting multiple columns

Several scenarios may occur where you would need to delete unnecessary columns from your datasets, in order to remove empty columns, or delete information that is not relevant anymore.






Talend Data Preparation allows you to delete multiple columns in one single action.

Let's take the example of a dataset containing some customers data, but where the information on the age, salary or marital status is missing. This leaves you with three empty columns, that you will delete in one go, with the dedicated function, to clean your dataset.

	First_Name first_name	Last_Name text	Gender gender	Age	Occupation text	Salary_Out text	MaritalStatus text	Address address_line	City city
1	Janes	Butt	F		K-12 Student			6649 N Blue Gum St	New Orleans
2	Josephine	Darakjy	M		Self-Employed			4 B Blue Ridge Blvd	Brighton
3	ART	Venere	M		Scientist			8 W Cerritos Ave #54	Bridgeport
4	Lenna	Paprocki	M		Executive/Managerial			639 Main St	Anchorage
5	Donette	Foller	M		Writer			34 Center St	Hamilton
6	Simona	Morasca	F		Homemaker			3 McAuley Dr	Ashland
7	Mitsue	Tollner	M		Academic/Educator			7 Eads St	Chicago
8	Leota	Billiard	M		Programmer			7 W Jackson Blvd	San Jose
9	Sage	Wieser	M		Technical/Engineer			5 Boston Ave #88	Sioux Falls
10	Kris	Marrier	F		Academic/Educator			228 Runamuck Pl #280	Baltimore
11	minna	Amigon	F		Academic/Educator			2371 Jerrold Ave	Kulpsville

1. While pressing the **Ctrl** key, click the headers of every column you want to delete, **Age**, **Salary_Out** and **MaritalStatus** in this example.

All three columns are now selected and highlighted in blue.

Gender 	Age 	Occupation 	Salary_Out 	MaritalStatus 
gender	text	text	text	text
F		K-12 Student		
M		Self-Employed		
M		Scientist		
M		Executive/Managerial		
M		Writer		

2. In the **Functions** panel, select the **Delete column** function.

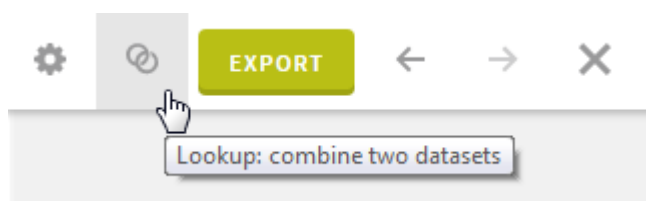
The function directly applies and the three columns are deleted. Three recipe steps are created, one for each column.



It is also possible to delete an individual column by clicking the menu icon in a column header and selecting **Delete Column**. However, using the dedicated function is the most efficient way to delete multiple columns at the same time.

Dynamically using the data from another dataset

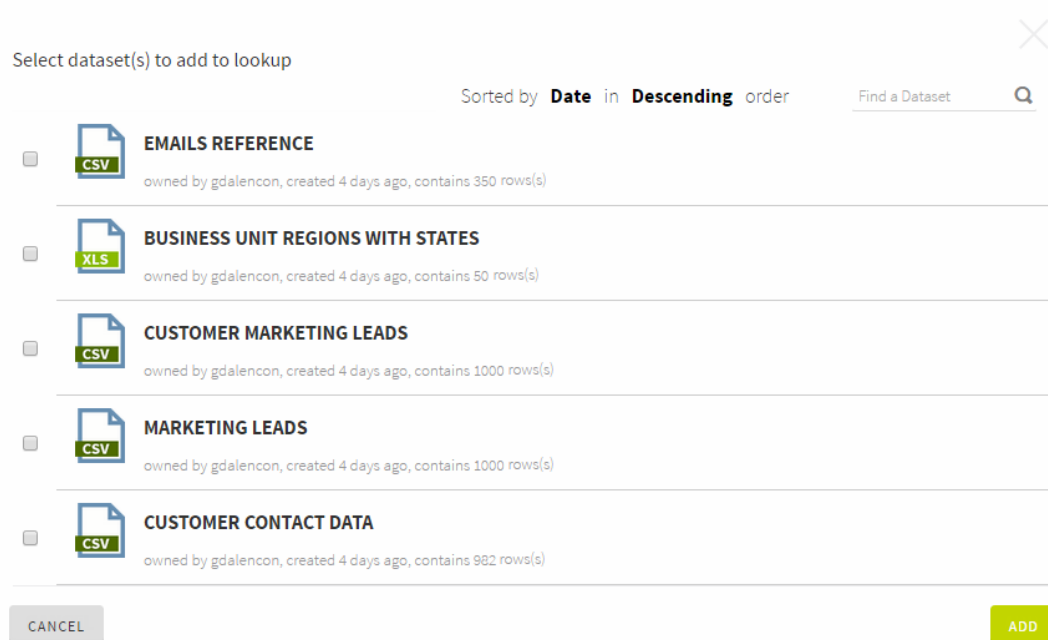
The lookup feature matches data from the current dataset with its counterpart in a "reference" dataset. For example, you can use it to add the full name of a US State alongside its abbreviation.

1. Select the column on which you want to perform the lookup.
This will be the source column for your data.
2. Click the lookup button to open the lookup panel.



3.  Click the  button and, in the dialog box that opens, select the dataset you want to use to perform the lookup.

This dataset contains the target column for your data.



4. Select the **Add to Dataset** check box under every column you want to include in your lookup.

	COMPANY_NAME text	EMAIL_DOMAIN web_domain
		<input checked="" type="checkbox"/> Add to Dataset
1	Abata	abata.com
2	Abatz	abatz.info
3	Agimba	agimba.biz

5. Point your mouse over the **Confirm** button to preview the changes.
6. Click the **Confirm** button to apply those changes.

A new column is created with the result of the lookup. Whenever the column in the lookup dataset appears in the main dataset, the associated column in the lookup dataset is added to the main dataset. If a row in the linked columns matches between the two datasets, the content of the associated column for this row is added too.

Extracting email address parts

An email address, such as *user@talend.com*, is made up of two parts separated by the @ symbol: the local part (*user* in this example) and the domain part (*talend.com* in this example).

The two parts of an email address can be extracted and copied to two new columns.

1. Select a column containing email addresses.
2. In the **Functions** panel, type **Extract Email Parts** and point your mouse over the function name to preview the result of the **Extract Email Parts** function.

EMAIL	EMAIL_LOCAL	EMAIL_DOMAIN	SU
email	text	text	
jbutt@gmail.com	jbutt	gmail.com	
josephine_darakjy@darakjy.org	josephine_darakjy	darakjy.org	
art@venere	art	venere	
lpaprocki@hotmail.com	lpaprocki	hotmail.com	
donette.foller@cox.net	donette.foller	cox.net	
simona@morasca.com	simona	morasca.com	
mitsue_tollner@yahoo.com	mitsue_tollner	yahoo.com	

Find a function ...

- SPLIT
- Extract email parts
- Extract number
- Extract string parts...
- Extract URL parts
- Split the text in parts...

3. Click the result to execute the **Extract Email Parts** function.

The local part and the domain part are extracted from the email addresses. The extracted data is put into two new columns.

Extracting part of a text

If you want to take part of the text (substring), contained in a cell and reuse it elsewhere, you can extract part of the text.

It is recommended to [remove unnecessary blank spaces from the text records](#) and to make sure the cells have the same length before proceeding.

1. Select a column from which you want to extract some text.
2. In the **Functions panel**, type **Extract Parts** of the **Text** and click the result to display the options of the associated function.
3. Choose the text you want to extract using the options.

Index represents a number of characters from which or to which the selection applies.

Extract parts of the text...

From:

From index

Beginning index:

17

To:

To N before end

To N before end:

2

SUBMIT

4. Click the **Submit** button to extract the selection you made to a new column.

The text corresponding to the selection you made is extracted to a new column.

Extracting the numbers contained in a column

If you want to use the numbers contained in a column, such as the postal code in an address, you may want to extract this numbers to a new column.

1. Select the column from which you want to extract numbers.
2. In the **Functions panel**, type `Extract Number` and click the result to execute the associated function.

The numbers contained in the selected column are extracted to a new column. If the numbers are written using metric prefixes, such as 1k instead of 1000, they will be extracted and written using the American standard.





Filling empty cells with text




The white part of the **Quality Bar** indicates that the column contains empty records.

You may want to fill these empty records with some text or with the content of another column.

1. Select a column containing empty records.

You can identify columns containing empty records using the **Quality Bar**. The white part of it indicates that the column contains empty records.

ISCUSTOMER?	REVENUE	COMPANY_NAME
boolean	decimal	text
	 	

-  **Green** – Valid data that matches the column type
-  **White** – Empty cells
-  **Orange** – Invalid data that does not match the column type

2. In the **Functions panel**, type `Fill Empty Cells with Text` and click the result to open the options for the associated function.

3. Click the **Submit** button to apply the function.

This fills all the empty cells in the selected column with text or data from another column.

Identifying if a column matches a pattern

If you want to identify whether a column matches a pattern you define, with or without a regular expression, you can use the **Matches Pattern** function.

1. Select the column on which you want to apply the pattern.
2. In the **Functions panel**, type `Matches Pattern` and click the result to open the options for the associated function.

3. Click the **Submit** button to apply the function.

This creates a new column with the value **true** if the pattern matches and **false** if it does not.

Merging the content of two or more columns

In some cases, the data you want to use is split in several columns. You can group these columns using a concatenation.

1. Select the column you want to use for the concatenation. This column will be the first part of the merged column that will be created.
2. In the **Functions panel**, type `Concatenate with` and click the result to display the options of the associated function.

3. Configure the concatenation options corresponding to your needs.
4. Click the **Submit** button to apply the function.

The content of two columns is merged using a concatenation.

Masking data

When manipulating sensitive data, such as names, addresses, credit card or social security numbers, you might want to mask this data.

To protect the original data while having a functional substitute, you can use the **Mask data (obfuscation)** function.

1. Select the column on which you want to apply the data masking.
2. In the **Functions panel**, type `Mask data (obfuscation)` and click the result to execute the associated function.

The data in the column has been replaced by random but usable substitutes. Depending on the semantic type of the column on which you use the **Mask data (obfuscation)** function, the effect will vary. For more information, see [Data Masking](#) on page 118



Putting the first letter of every word in upper case

If you want the first letter of every word to appear in upper case, you can use the **Change style to Title case** function.

1. Select a column containing text for which you want every word to start with an upper case letter.

LAST_NAME	text
butt	
darakjy	
venere	
paprocki	

2. In the **Functions** panel, type `Change to title case` and click the result to execute the associated function.

COLUMN	ROW
change to title case	
SUGGESTIONS	
Change to title case	

3. Repeat this action for every column you want to modify.

Every word starts with a capital letter in the selected column.

LAST_NAME	text
Butt	
Darakjy	
Venere	
Paprocki	

Removing unnecessary blank spaces in text records

Blank spaces can be present before and after the content from each cell.

They are more likely to be present in columns containing data manually entered by someone, such as a name or a phone number. These spaces are shown as grey squares.

1. Select a column containing text with blank spaces or a column that you think may contain some.

5	jagc	mic
10	kris	Mari
11	afac	afac

2. In the functions panel, type `Remove trailing and leading characters` and click the result to open the options the associated function.

3. In the **Padding character** drop-down list, select **whitespace**.

Select **other** to specify any other trailing or leading character to be removed.

4. Click **Submit**.

Blank spaces are removed from the selected column.

Removing all the empty and invalid rows

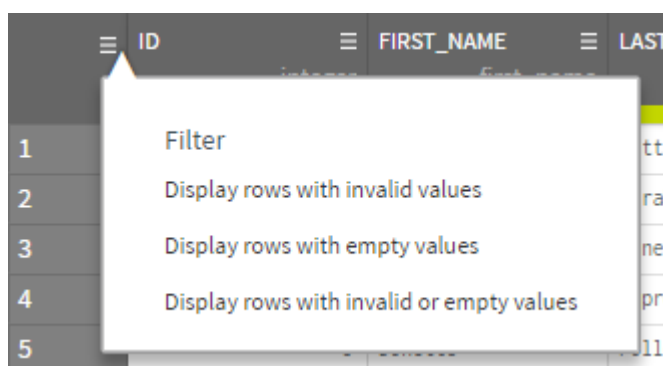
Using the quality bar is a convenient way to filter and remove invalid rows for a given column, but this action is also available for the whole dataset.

You can apply a filter on all the invalid and empty rows from your dataset to remove them in a single action.

Let's take the example of a dataset containing customer data, where some phone numbers and email addresses entries are either invalid or empty.

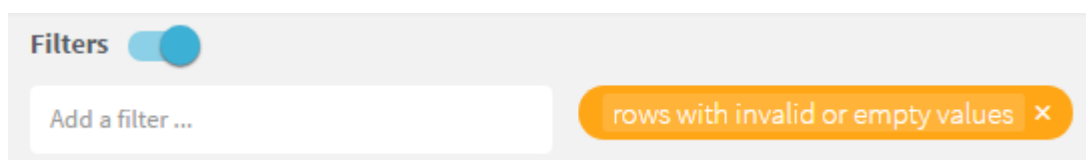
PHONE	EMAIL
us_phone	email
504-621-8927	jbutt@gmail.com
810-292-9388	josephine_darakjy@da
856-636-	art@venere
	lpaprocki@hotmail.com

1. Click the white arrow on the top left of the grid.



2. Select **Display rows with invalid or empty values**.

The filter has been applied and the grid now only displays the rows containing at least one empty or invalid entries.



You can also choose to only filter the invalid or empty rows to remove them from your dataset.

3. In the functions panel, search the **Delete these Filtered Rows** function and click it to apply it on you data.

The rows have been deleted and you can now remove the filter.

4. In the filter bar, click the cross in the filter or click the garbage bin icon to display the whole dataset again.

Your dataset is now free of any invalid or empty values and the quality bar is fully green for all the columns.

PHONE	EMAIL
us_phone	email
504-621-8927	jbutt@gmail.com
810-292-9388	josephine_darakjy@da
419-503-2484	simona@morasca.com
605-414-2147	sage_wieser@cox.net

Reordering preparation steps

In Talend Data Preparation, each preparation step you apply on your data is based on the previous one. As a consequence, if you already applied many preparation steps to your data, but forgot one small change at the beginning, you would not achieve the expected result.

In a preparation with many steps, you have the possibility to rearrange your preparation steps so that the changes take effect in the right order.

Let's take the example of a dataset containing customer information such as their name, email, adress or the US State they live in.

	ID	FIRST_NAME	LAST_NAME	GENDER	AGE	CITY	STATE	EMAIL
	integer	first_name	text	gender	text	city	us_state_code	email
1	1	James	Butt	F	Under 18	New Orleans	LA	jbutt@gmail.com
2	2	Josephine	Darakjy	M	56+	Brighton	MI	josephine_darakjy@darakjy.oi
3	3	Cammy	Albares	M	25-34	Laredo	Texas	calbares@gmail.com
4	4	Lenna	Paprocki	M	45-49	Anchorage	AK	lpaprocki@hotmail.com
5	5	Donette	Foller	M	25-34	Hamilton	OH	donette.foller@cox.net
6	6	Simona	Morasca	F	50-55	Ashland	OH	simona@morasca.com
7	7	Mitsue	Tollner	M	35-44	Chicago	IL	mitsue_tollner@yahoo.com

Based on this dataset, a few preparation steps have already been made, including a lookup on the **State** column, some cleansing actions, and lastly, a correction on the **State** column, where one of the States was in the wrong format.

Reordering steps

1

Lookup done with dataset STATES. Join has been set between STATE and STATE. The column REGION has been added.

2

Delete the rows with invalid cell on column EMAIL

3

Change to upper case on column LAST_NAME

4

Replace the cells that match on column STATE

Current:

= Texas

Replacement:

TX

☐ Overwrite entire cell

SUBMIT

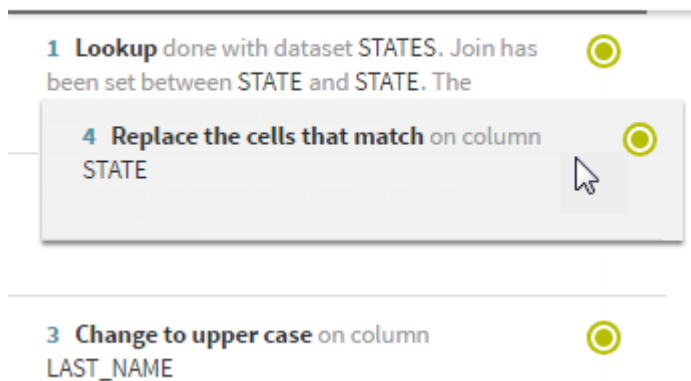
Because that last change on the **State** column was performed after the lookup, some information is now missing from the **Region** column.

STATE	REGION
us_state_code	city
LA	South East
MI	Mid West
TX	

You are going to take the lookup step and place it as the final step of the preparation to make sure it includes all the States.

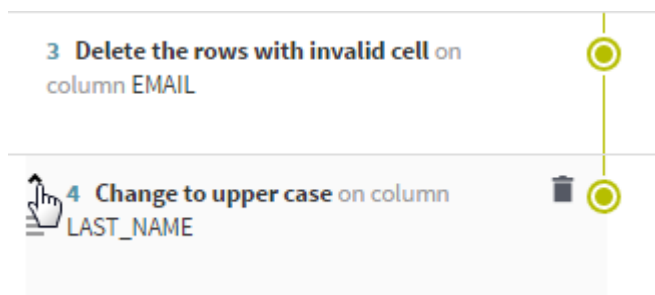
1. Point your mouse over the lookup step.
2. To move the lookup step from the fourth position to the first position, you can:
 - Either drag the recipe step and drop it at the top of your recipe.

Reordering steps



The grey line shows where the recipe step will be placed.

- Or click the up arrow on the left of your recipe step to move it up one step at a time.



Your preparation is automatically updated with the correct sequence of actions, and the **Region** column now includes Texas.

	ID	FIRST_NAME	LAST_NAME	GENDER	AGE	CITY	STATE	REGION	EMAIL
	integer	first_name	text	gender	text	city	us_state_code	city	email
1	1	James	BUTT	F	Under 18	New Orleans	LA	South East	jbutt@gmail.com
2	2	Josephine	DARAKJY	M	56+	Brighton	MI	Mid West	josephine_darakjy@darakjy.oi
3	3	Cammy	ALBARES	M	25-34	Laredo	TX	South West	calbares@gmail.com
4	4	Lenna	PAPROCKI	M	45-49	Anchorage	AK	West	lpaprocki@hotmail.com
5	5	Donette	FOLLER	M	25-34	Hamilton	OH	Mid West	donette.foller@cox.net
6	6	Simona	MORASCA	F	50-55	Ashland	OH	Mid West	simona@morasca.com
7	7	Mitsue	TOLLNER	M	35-44	Chicago	IL	Mid West	mitsue_tollner@yahoo.com

Replacing the content of cells with another value

If you want to replace the content of a column with a given value or with the content of another column, you can use the **Fill Cells with Value** function.

1. Select a column in which you want to replace values.
2. In the **Functions** panel, type **Fill Cells with Value** and click the result to open the options for the associated function.

3. Click the **Submit** button to apply the function.

The content of the selected column is replaced with the data you selected.

Using regular expressions to match content

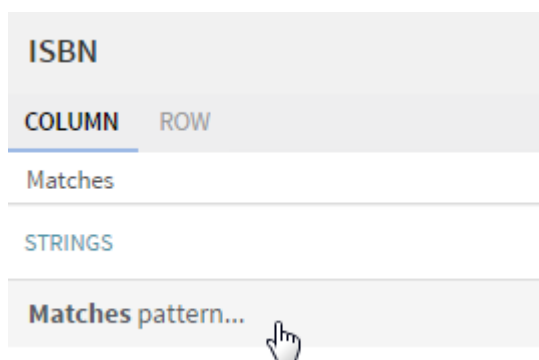
Regular expressions can be used to search for a specific pattern among your data and isolate values that you are interested in.

This scenario takes the example of someone working on a dataset that lists information about books, including their ISBN numbers. Using Talend Data Preparation, it is possible to check if the ISBN are valid, and follow the right pattern. With the **Matches Pattern** function, you can compare your data with an expression of your choice.

1. Click the **ISBN** column to select its content.

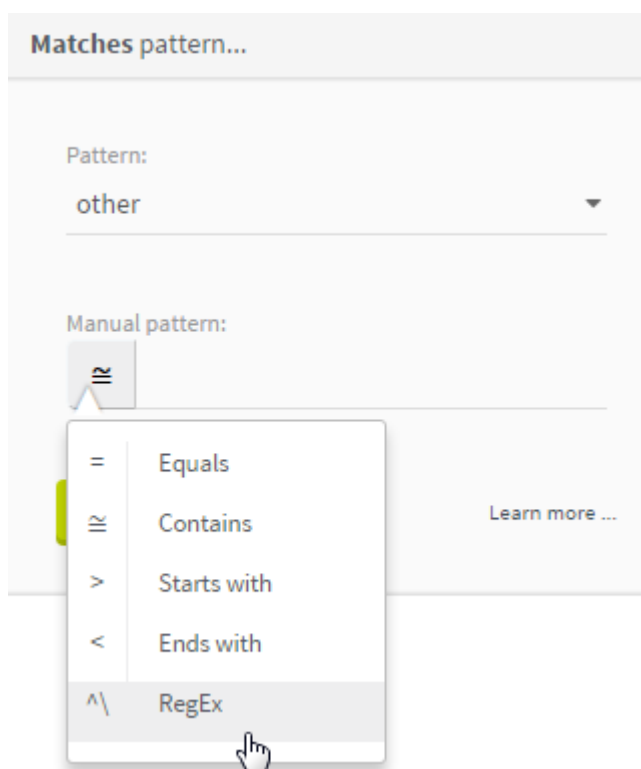
ISBN
ISBN 713861304-0
992424254-8
ISBN 2-226-05257-7
ISBN 2-277-30048-9
386236291-4
ISBN 2-080-72115-1
151345594-X

2. In the functions list, find and select **Matches Pattern...**



A menu opens where you can enter the pattern for your search.

3. In the **Pattern** field, select **other** from the drop-down list.
4. Click the button on the left side of the **Manual pattern** field and select **RegEx** from the list.



5. In the **Manual pattern** field, type `^[ISBN]{4}[]{0,1}[0-9]{1}[-]{1}[0-9]{3}[-]{1}[1]{0-9}{5}[-]{1}[1]{0-9}{0,1}$`.

This regular expression corresponds to the ISBN number model that you want to identify in your dataset.

6. Click **Submit**.

A new column **ISBN_MATCHING** is created, where the values that match the pattern defined by the regular expression, are listed as **true**. The values that do not match are listed as **false**.

ISBN ≡ text	ISBN_MATCHING ≡ boolean
ISBN 713861304-0	false
992424254-8	false
ISBN 2-226-05257-7	true
ISBN 2-277-30048-9	true
386236291-4	false
ISBN 2-080-72115-1	true
151345594-X	false

After using a regular expression to search for a specific pattern, you can now easily identify and isolate the values that match your search.

Working on large datasets

By default, a dataset that exceeds 10,000 rows for Talend Data Preparation, and 30,000 rows for Talend Data Preparation Free Desktop is considered a large dataset.

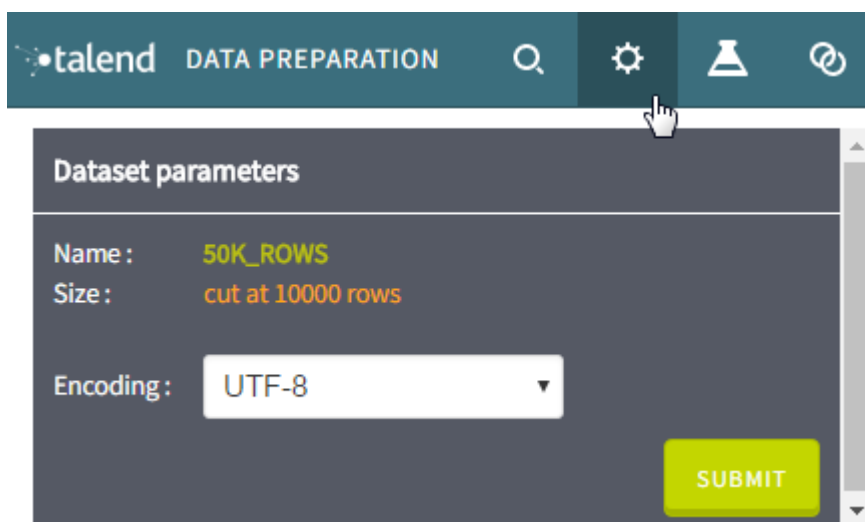
Even if there is no limitation regarding the size of the dataset that you can import, the export settings and the display of large datasets are different than usual. Let's take the example of a dataset containing 50,000 rows:

- In Talend Data Preparation Free Desktop, the import will be cut at 30,000 rows. You can only prepare and export the first 30,000 rows of your dataset. This is a default value that can be set to a lower value by editing the `dataset.records.limit` parameter in the `application.properties` file, located in the installation folder.
- In Talend Data Preparation, you will be able to work on a sample displaying the first 10,000 rows. This is a default value. You can set it to a higher value by editing the `dataset.records.limit` parameter in the `application.properties` file, located in the installation folder. A higher maximum value might decrease the application performances. The maximum value that you can set depends on your Web browser, your network quality, and the power of your machine. Do not exceed 100,000 rows as maximum value for your sample.

If you change the default value for the number of rows to be displayed, it will only apply to datasets that are imported from this point onwards, and not to existing datasets.

Fetching more data from a large dataset

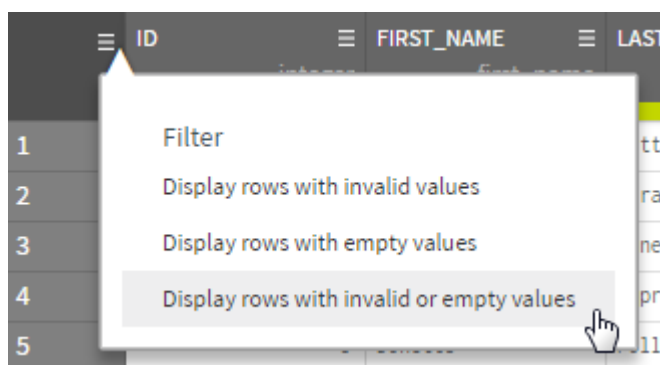
When working on a large dataset in Talend Data Preparation, 50,000 rows for example, only a sample of the first 10,000 rows is displayed, as you can see in the dataset parameters:



You can start preparing your data and apply functions, like you would normally do for any other dataset.

One difference occurs when you apply a filter of any type on your data. Since you are working on a sample, only the matching rows among the first 10,000 will be retrieved. But you have the possibility to fetch more matching rows, among the remaining 40,000 and refine your preparation based on this new sample.

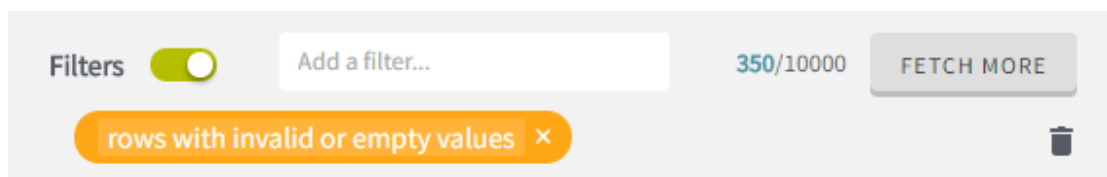
1. Click the white arrow on the top left of the grid and select **Display rows with invalid or empty values**.



You can see in the filter bar that the filter has been correctly applied and only the matching rows are displayed in the grid. You can choose any other filter.

You can also notice the **Fetch more** button in the filter bar, showing that you are currently working on a sample, and that more rows potentially match your filter.

2. Click **Fetch more rows**, to retrieve more rows matching your current filters.



The **Fetch additional rows** dialog box opens, where you can see the status of the data retrieval.

FETCH ADDITIONAL ROWS

Fetching additional rows

Currently fetching more rows matching the filters you have defined.
 Rows fetched: **685/10000**.
 Wait for Talend Data Preparation to gather up to rows, or click "Show fetched rows" to interrupt the process and display the rows already fetched.

CANCEL

SHOW FETCHED ROWS

Talend Data Preparation automatically stops when it reaches 10,000 results, or the end of the dataset. You also have the possibility to stop the process and show the rows already found. You are then taken back to the grid, where the fetched rows now form the sample you will be working on. Any filter or function applied from now on will only apply to this sample.

If the filter you initially chose to apply doesn't match any row, you can either clear all your filters, or try and search the whole dataset for matching rows.

▼	CITY	▼	STATE	▼	REGION
text	city	us_state_code			

No rows matching your filter in the sample:
 You can click [here](#) to look into the entire dataset
 or
 You can click [here](#) to remove all your filters.

3. To go back to your initial sample, clear all your filters.

Click the cross in each individual filter or click the garbage bin icon to clear the filters.

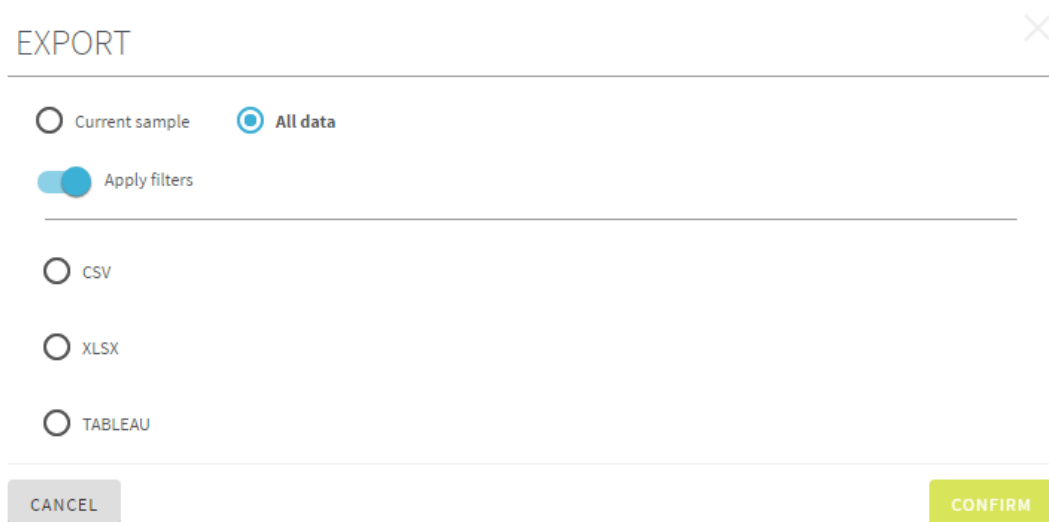
The grid now displays the first 10,000 rows of your dataset again and you can continue preparing your data.

Exporting a preparation made on a large dataset

When you are finished preparing a large dataset, you have the choice to export only the sample you were working on, or the full prepared data.

If you still had filters applied on your data at the moment of the export, you can also decide if you want to keep them activated or not.

1. Click the **Export** button, on the top right of the grid



The image shows a dialog box titled "EXPORT" with a close button (X) in the top right corner. Inside the dialog, there are two radio buttons: "Current sample" and "All data". The "All data" radio button is selected. Below these, there is a toggle switch labeled "Apply filters" which is currently turned on. Underneath the toggle, there are three radio buttons for the output format: "CSV", "XLSX", and "TABLEAU". At the bottom of the dialog, there are two buttons: "CANCEL" on the left and "CONFIRM" on the right.

2. Select your export option:

- If you select **All data**, all the preparations steps you have performed on the 10,000 row sample will be applied to the rest of the dataset as well.
- If you select **Current sample**, only the first 10,000 rows by default will be exported.

3. Choose a filename for the export and select the output format between CSV, XLSX or Tableau.

4. Click Confirm.

- If you choose to export only the sample, your download of the output file directly starts.
- If you choose to export the full data, the export process is launched in the background. You can check the status of the export, and download your output file in the **Export history** page. For more information, see [The export history page](#) on page 150.

5. When the export is ready, click the Download button to retrieve the result of your preparation.

The preparation steps have been applied to the whole dataset, and you can now find the result of your preparation in your Downloads folder.

If you have already exported a file from this preparation, you can choose to keep your previous export settings, or make a **New Export** with new settings.

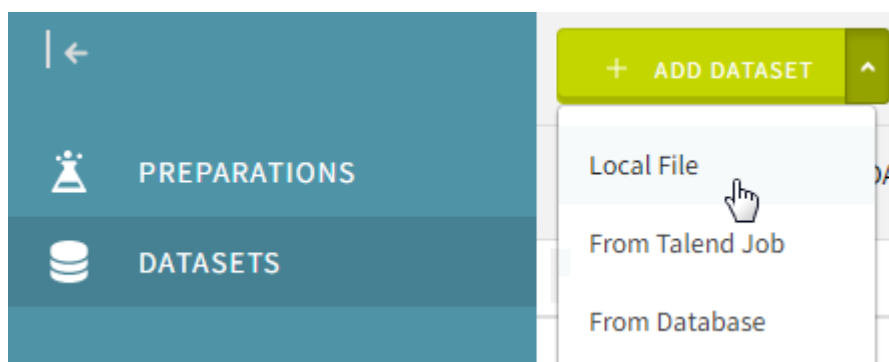
Exporting again with the same settings will also overwrite your last export set to **All data**.

Adding a dataset from a local file

A dataset holds the raw data that can be used as the raw material for one or more preparations.

The easiest way to import data into the application is to create a dataset from one of your local files.

- 1. In the Datasets view of the Talend Data Preparation homepage, click the white arrow next to the Add Dataset button.**
- 2. Select Local File.**



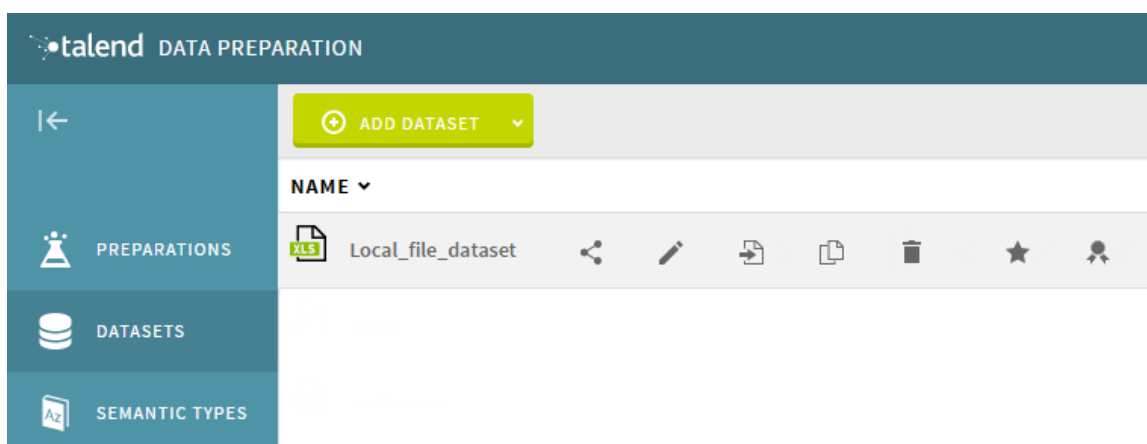
3. Browse your files to select the one to import.

You can import the following file types to use as dataset:

- XLS/XLSX
- CSV

Your dataset automatically opens and you can start your preparation.

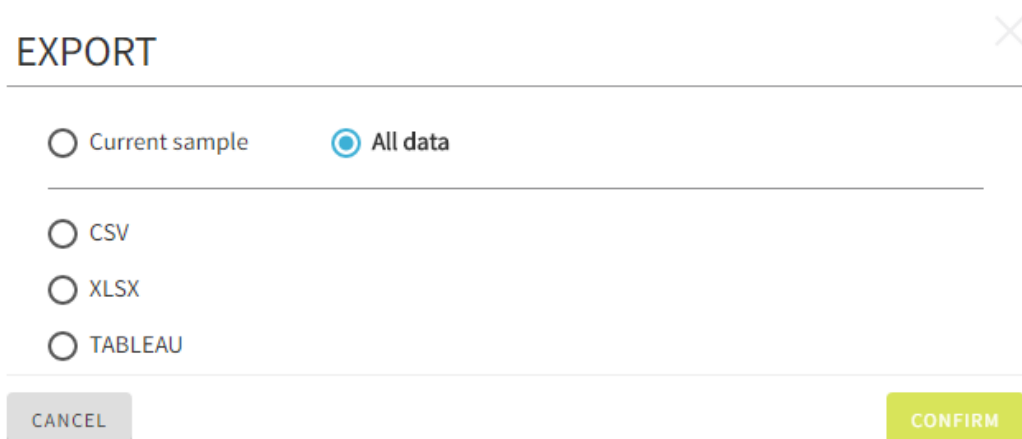
The dataset is added to the list in the **Datasets** view of the homepage.



Exporting a preparation made on a local file

When you are finished preparing your dataset based on a local file, you may want to export the data you have cleansed.

1. Click the **Export** button in the application header bar.



The image shows a dialog box titled "EXPORT" with a close button (X) in the top right corner. Inside the dialog, there are two radio buttons: "Current sample" and "All data". The "All data" radio button is selected. Below these, there are three more radio buttons: "CSV", "XLSX", and "TABLEAU". At the bottom of the dialog, there are two buttons: "CANCEL" on the left and "CONFIRM" on the right.

2. Choose the file format you want to use when exporting your data.

- If you choose **CSV**, choose a delimiter to use and enter a name for the file to export.
- If you choose **XLSX**, or **Tableau** choose a name for the file to export.

3. Click **Confirm**.

The export operation is processed on the Talend Data Preparation server.

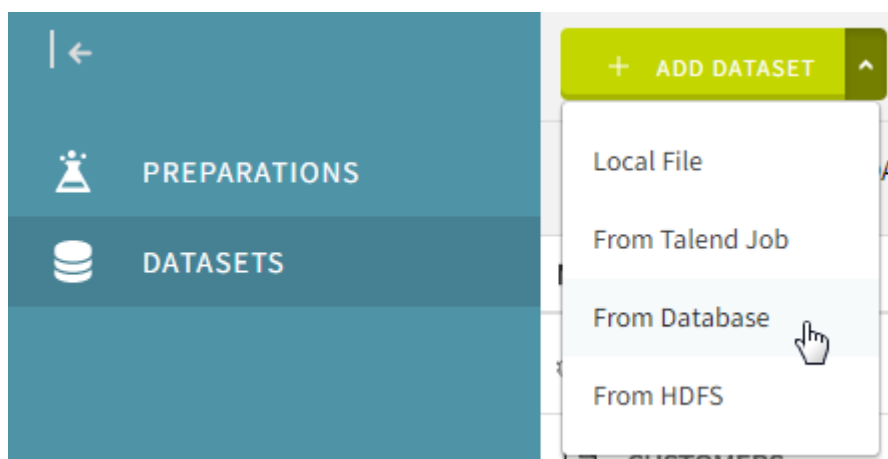
If the result of your preparation is larger than your current sample size, 10 000 rows by default, you can choose between exporting just the sample, or the whole data. In the first case, your download of the output file directly starts. In the second case, the export process is launched in the background. You can check the status of the export, and download your output file in the **Export history** page. For more information, see [The export history page](#) on page 150.

Adding a dataset from a database

Talend Data Preparation is able to connect to various databases and use them as source to create a new dataset.

In this example, you want to prepare some customers data that is stored on a MySQL database. You will enter your database connection information, directly in the Talend Data Preparation interface and create a new dataset from this data.

1. In the **Datasets** view of the Talend Data Preparation homepage, click the white arrow next to the **Add Dataset** button.
2. Select **From Database**.



The **Add a database dataset** form opens.

3. In the **Dataset name** field, enter the name you want to give your dataset.
4. In the **Database type** drop-down list, select the type of database you want to connect to, **MySQL** in this example.

This list can be manually enriched. For more information, see [Adding a new database type](#) on page 90.

5. In the **JDBC URL** field, provide a URL to access your MySQL database.

The form provides a URL template where you can adapt the values to match your own connection details:

- Replace `localhost` with your IP address.
- Replace `3306` with the port that you have set for MySQL. 3306 is the default port for MySQL.
- Replace `db` with the name of the database you want to connect to.

6. In the **Username** and **Password** fields, enter your MySQL connection information.
7. Click **Test connection**.

If the connection is successful, the second part of the form is displayed, where you can enter a query for your database. If not, an error message is displayed, detailing why the connection failed.

Make sure that MySQL authorizes connection from Talend Data Preparation.

8. In the **Query** field, enter the query for the information that you want to retrieve from the table stored in your database.

Query

`select * from customers`

CANCEL ADD DATASET

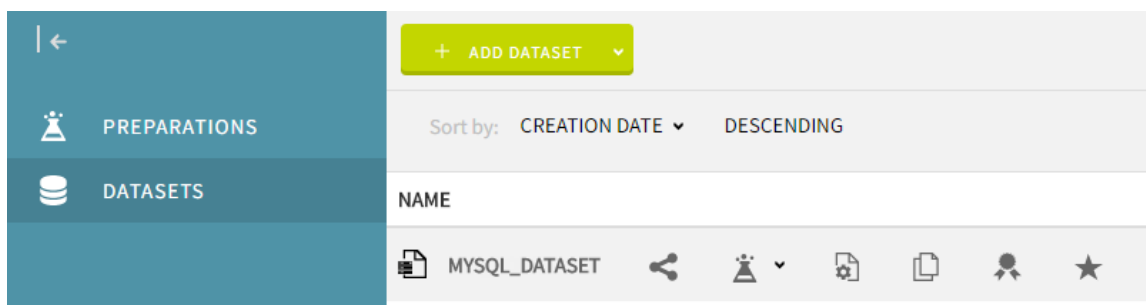
In this case, all the information from the table called `customers` will be retrieved and output as a dataset.

9. Click the **Add dataset** button at the end of the form.

The data extracted from the `customers` table in your MySQL database directly opens in the grid and you can start working on your preparation the same way you usually do.

The data is still stored in the MySQL database, Talend Data Preparation only retrieves a sample on-demand.

The dataset is added to the list in the **Datasets** view of the homepage.



Adding a new database type

Talend Data Preparation allows a direct connection to various types of databases. You can use them as source to create new datasets. By default, Talend Data Preparation offers connectivity to MySQL, Derby, PostgreSQL, SQL Server and Azure SQL databases.

It is possible to manually enrich the list of databases from which you can import data.

The list of available database types for dataset creation actually depends on the JDBC drivers that you have stored in the `<components_catalog_path>/ .m2` folder.

Let's say that you have some customer data stored on an Oracle database, and you want to import it in Talend Data Preparation to perform cleansing operations. You will add a JDBC driver `.jar` file specific to Oracle databases to the Components Catalog folder structure to add this new source of data in the Talend Data Preparation interface.

In a Big Data context, if you want to run preparations made on data from your Oracle database, on the Hadoop cluster, the same driver must be added to the Spark Job Server folder structure.

You do not need to stop or restart any of the services to complete the following procedure.

The Components Catalog server and the Spark Job Server are installed and running on a Linux machine.

1. Download the latest Oracle jdbc driver called `ojdbc7.jar` from the [Oracle website](#).
2. Create the `<components_catalog_path>/m2/jdbc-drivers/oracle/7/` folder.



Warning: The folder structure must follow this template: `.m2/jdbc-drivers/<database_name>/<jdbc_version>`.

3. Copy the `ojdbc7.jar` in the newly created folder.
4. Change the name of the file from `ojdbc7.jar` to `oracle-7.jar`.



Warning: The file name must follow this template: `<database_name>-<jdbc_version>`.

The purpose of renaming the `.jar` file and the folder structure, is to ensure naming consistency and make them Maven compliant.

5. Update the `<components_catalog_path>/config/jdbc_config.json` file by adding the following lines:

```
,
{
  "id" : "Oracle Thin",
  "class" : "oracle.jdbc.driver.OracleDriver",
  "url" : "jdbc:oracle:thin:@myhost:1521:thedb",
  "paths" :
  [
    { "path" : "mvn:jdbc-drivers/oracle/7" }
  ]
}
```

Where:

- `id` is the value that will be displayed in the Talend Data Preparation interface as **Database type**.
 - `class` is the driver class used to communicate with the database.
 - `url` is the URL template to access a database.
 - `path` follows this model: `mvn:jdbc-drivers/my_database_name/my_version`
6. To enable export on the Hadoop cluster for the new dataset type, copy the `oracle-7.jar` file to the `<spark_job_server_path>/datastreams-deps/` folder.
 7. Copy the changes made in the `<components_catalog_path>/config/jdbc_config.json` file, and paste them into the `<spark_job_server_path>/jdbc_config.json` file.

The Oracle database is now available in the **database type** drop-down list in the import form.

When exporting a preparation made on data stored on your Oracle database, you can choose to process the data on the Talend Data Preparation server, or a Hadoop Cluster if you are using Big Data.

For more information on how to import data from a database, see [Adding a dataset from a database](#) on page 88.

Exporting a preparation made on a database dataset

When you are finished preparing a dataset extracted from a database, you may want to export your data.

1. Click the **Export** button in the application header bar.

EXPORT TO HDFS

☐ Current sample
 ☒ All data

☐ CSV
☐ XLSX
☐ TABLEAU
☒ HDFS

Format:
CSV

Delimiter:
☒ Semicolon
 ☐ Tab
 ☐ Space
 ☐ Comma
 ☐ Pipe

Output path:
<file_path_on_cluster>

Authentication method:
Specified kerberos

Keytab:
/keytabs/mykeytab.keytab

Principal:
user@realm.com

CANCEL CONFIRM

2. If the result of your preparation is larger than your current sample size, 10 000 rows by default, select an export option:
 - If you select **Current sample**, only the sample you have been working on will be exported.
 - If you select **All data**, all the preparations steps you have performed on your sample will be applied to the rest of the dataset as well.
3. Choose between exporting your data to a local file, or to a Hadoop cluster
 - If you export your data as a `csv` or `xlsx` local file, the export operation will be processed on the Talend Data Preparation server.

- If you export your data to the Hadoop cluster, the export operation will be processed directly on the cluster. Choose the type of your output file between `csv`, `avro` or `parquet`. Enter the path to your preferred location on the cluster to save your file, and if you choose to authenticate via Kerberos, enter your principal and the path to your `keytab` file.

4. Click **Confirm**.

In the case of an export to a local file, if you chose to export only the **Current sample**, the download automatically starts. But if you selected **All data** to export the entire data, the export process is launched in the background. You can check the status of the export, and download your output file in the **Export history** page. For more information, see [The export history page](#) on page 150.

The export process triggers a refresh in the data that is fetched from the database, guaranteeing that the data displayed in the output is always up to date.

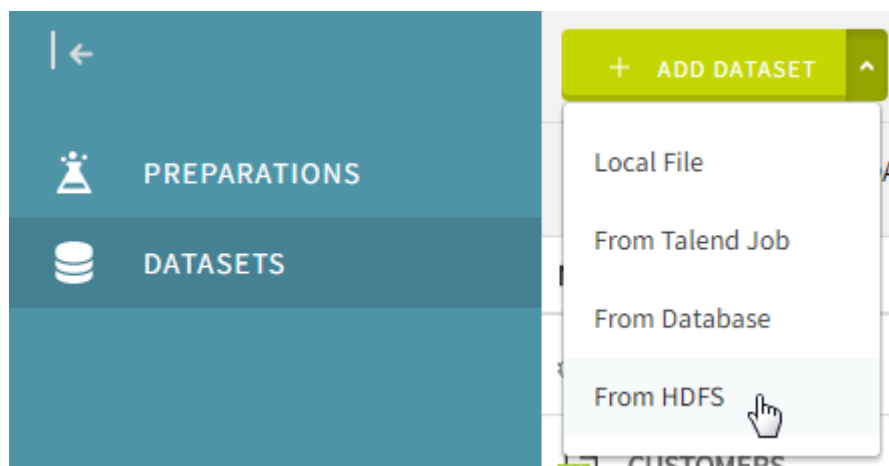
However, due to this refresh, it is possible that a dataset originally smaller than 10,000 rows, now exceeds this limit. In this case:

- If you export to a local file, only the sample is kept.
- If you export to a Hadoop cluster, the whole data is exported.

Adding a dataset from HDFS

You can access data stored on HDFS (Hadoop File System), directly from the Talend Data Preparation interface and import it in the form of a dataset.

1. In the **Datasets** view of the Talend Data Preparation homepage, click the white arrow next to the **Add Dataset** button.
2. Select **From HDFS**.



The **Add an HDFS dataset** form opens.

ADD A HDFS DATASET



Dataset name*
HDFS_dataset

☐ Use Kerberos

User name
Linux user

Format*
CSV

Path*
<file_path_on_cluster>

Record Delimiter
LF

Field Delimiter
SEMICOLON

CANCEL ADD DATASET

3. In the **Dataset name** field, enter the name you want to give your dataset.

4. In the **User name** field enter your Linux user name.

This user must have the reading rights on the file that you want to import.

5. To enable Kerberos authentication, select the **Use Kerberos** check box.

☒ Use Kerberos

Principal
username@example.com

Keytab file
/home/username/username.keytab

6. In the **Principal** field enter the name of the service principal.

7. In the **Keytab file** field, enter the location of your keytab file.

The keytab file must be accessible by the Spark Job Server.

You can manually configure Talend Data Preparation to display a default value in those fields.

8. In the **Format** field, select the format that corresponds to the file that you want to import.

For HDFS files, Talend Data Preparation supports CSV, AVRO and PARQUET.

If you choose **CSV**, select the record delimiter and field delimiter used for the file you want to import.

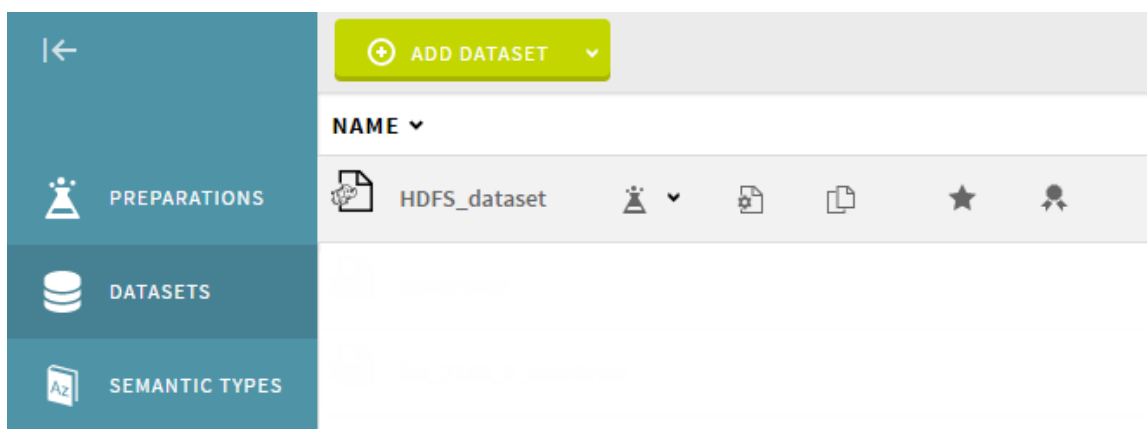
9. In the **Path** field, enter the complete URL of your file in the Hadoop cluster.

10. Click the **Add Dataset** button.

The data extracted from the cluster directly opens in the grid and you can start working on your preparation.

The data is still stored in the cluster and doesn't leave it, Talend Data Preparation only retrieves a sample on-demand.

Your dataset is now available in the **Datasets** view of the application home page.



Exporting preparation made on an HDFS dataset

When you are finished preparing your dataset extracted from HDFS, you have the possibility to export it back directly to the cluster, or download it as a local file.

Note that the cluster where you will export your cleansed data, must be the same cluster from which you imported the data in the first place.

1. Click the **Export** button in the application header bar.

EXPORT TO HDFS

☐ Current sample
 ☒ All data

☐ CSV
☐ XLSX
☐ TABLEAU
☒ HDFS

Format:
CSV

Delimiter:
☒ Semicolon
 ☐ Tab
 ☐ Space
 ☐ Comma
 ☐ Pipe

Output path:
<file_path_on_cluster>

Authentication method:
Specified kerberos

Keytab:
/keytabs/mykeytab.keytab

Principal:
user@realm.com

CANCEL CONFIRM

2. If the result of your preparation is larger than your current sample size, 10 000 rows by default, select an export option:

- If you select **Current sample**, only the sample you have been working on will be exported, as a local csv, xlsx or tableau file.
- If you select **All data**, all the preparations steps you have performed on your sample will be applied to the rest of the dataset as well, and the HDFS export will be enabled.

3. Select **HDFS**.

4. In the **Format** field, select the output format for your data.

For HDFS files, Talend Data Preparation supports CSV, AVRO and PARQUET.

If you choose **CSV**, select the delimiter to use for the output file.

5. In the **Path** field, enter the complete URL to your preferred location on the cluster to save the exported file.

6. If you chose to authenticate via Kerberos, enter your principal and the path to your keytab file.

The path must point to a keytab file that is accessible to all the workers on the cluster.

7. Click **Confirm**.

Note that if a preparation contains actions that only affect a single row, or cells, they will be skipped during the export process. The **Make as header** or **Delete Row** functions for example do not work in a Big Data context. A warning will be displayed before the export if your preparation contains such actions.

If you chose to export your sample as a local file, your download of the output file directly starts.

In the case of a full export, whether it is as a local file or to the cluster, the export operation starts in the background. You can check the status of the export, and download your output file in the **Export history** page. For more information, see [The export history page](#) on page 150.

The whole operation is processed directly on the Hadoop cluster.

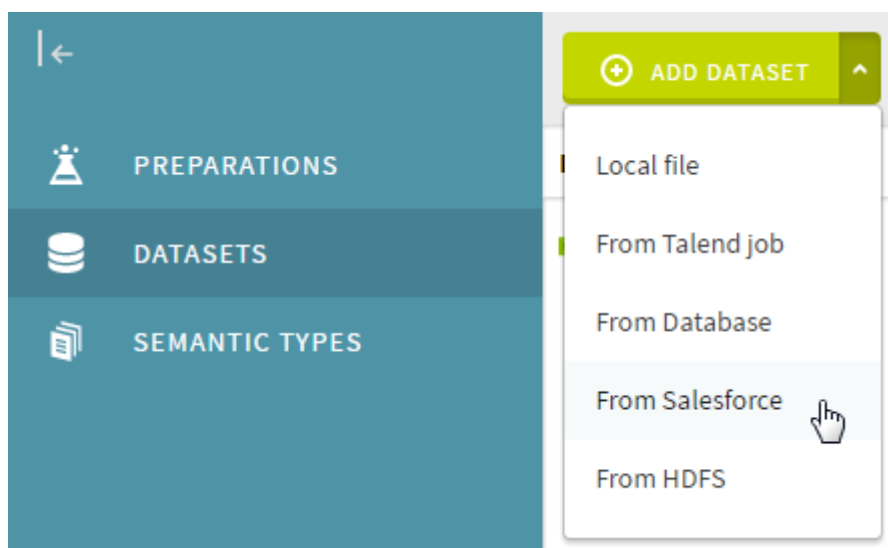
The export process triggers a refresh in the data that is fetched from the cluster, guaranteeing that the data displayed in the output is always up to date.

Adding a dataset from Salesforce

Talend Data Preparation is able to connect to various data sources to create new datasets.

In this example, you want to prepare some customers data that is stored on Salesforce. You will enter your Salesforce connection information, directly in the Talend Data Preparation interface and create a new dataset from this data.

1. In the **Datasets** view of the Talend Data Preparation homepage, click the white arrow next to the **Add Dataset** button.
2. Select **From Salesforce**.



The **Add a Salesforce dataset** form opens.

- If the connection is successful, the second part of the form is displayed, where you can enter a query or directly choose a Salesforce module from the list proposed. If not, an error message is displayed, detailing why the connection failed.

- Source type*

☒ Module selection ☐ SOQL query

Salesforce module*

AcceptedEventRelation

Account

AccountCleanInfo

AccountContactRole

AccountFeed

Talend Data Preparation User Guide (2017-06-29) | 98

Source type*
☒ Module selection ☐ SOQL query

Salesforce module*
 Account

Column Selection (5/51 selected) Q

<input type="checkbox"/>	All
<input checked="" type="checkbox"/>	Id
<input type="checkbox"/>	IsDeleted
<input type="checkbox"/>	MasterRecordId
<input checked="" type="checkbox"/>	Name
<input checked="" type="checkbox"/>	ParentId
<input type="checkbox"/>	BillingStreet
<input checked="" type="checkbox"/>	BillingCity
<input type="checkbox"/>	BillingState
<input type="checkbox"/>	BillingPostalCode
<input checked="" type="checkbox"/>	BillingCountry
<input type="checkbox"/>	BillingLatitude

CANCEL ADD DATASET

- Select the **SOQL query** radio button and in the **Query** field, enter the query for the information that you want to retrieve.

Source type*
☐ Module selection ☒ SOQL query

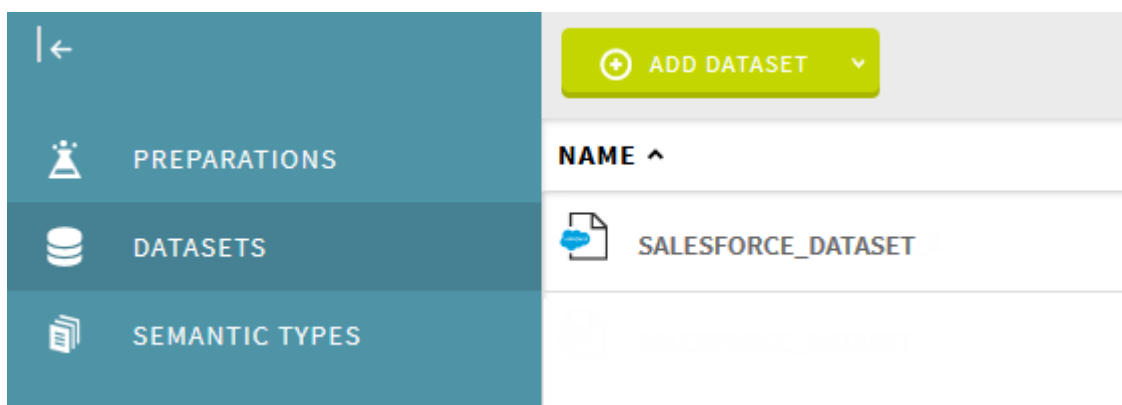
Query*
 select Id, Name, IsWon, FiscalYear, Account.Name FROM Opportunity

7. Click the **Add dataset** button at the end of the form.

The data extracted from Salesforce directly opens in the grid and you can start working on your preparation the same way you usually do.

The data is still stored in Salesforce, Talend Data Preparation only retrieves a sample on-demand.

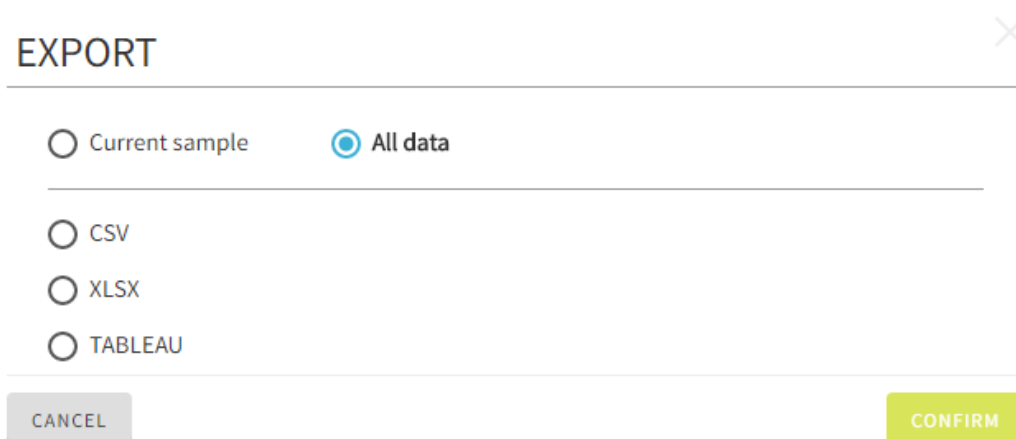
The dataset is added to the list in the **Datasets** view of the homepage.



Exporting a preparation made on a Salesforce dataset

When you are finished preparing a dataset extracted from Salesforce, you may want to export your data.

1. Click the **Export** button in the application header bar.



2. If the result of your preparation is larger than your current sample size, 10000 rows by default, select an export option:
 - If you select **Current sample**, only the sample you have been working on will be exported.
 - If you select **All data**, all the preparations steps you have performed on your sample will be applied to the rest of the dataset as well.
3. Choose the file format you want to use when exporting your data.
 - If you choose **CSV**, choose a delimiter to use and enter a name for the file to export.
 - If you choose **XLSX**, or **Tableau** choose a name for the file to export.
4. Click **Confirm**.

The export operation is processed on the Talend Data Preparation server.

If you chose to only export the **Current sample**, your download of the output file directly starts. If you chose **All data**, the export process is launched in the background. You can check the status of the export, and download your output file in the **Export history** page. For more information, see [The export history page](#) on page 150.

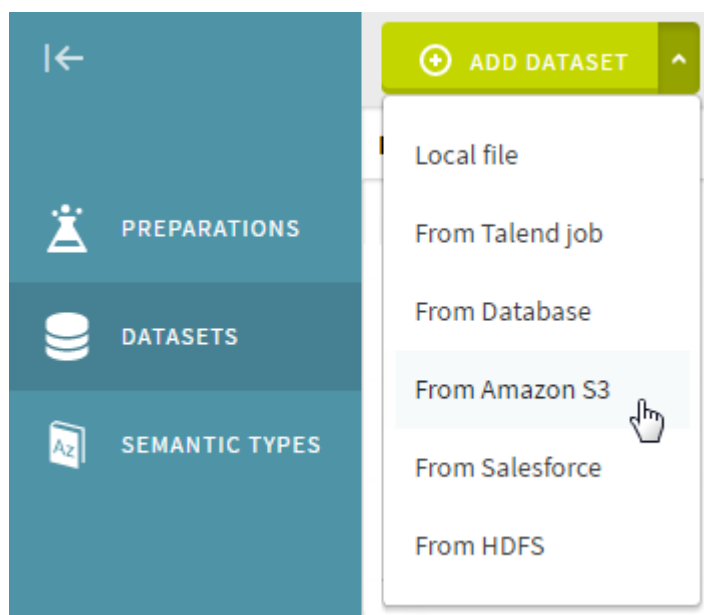
If you use Talend Data Preparation in a Big Data context, it is also possible to export the result of your preparation to a Hadoop File System.

Adding a dataset from Amazon S3

Talend Data Preparation is able to connect to various data sources to create new datasets.

In this example, you want to prepare some customers data that is stored on Amazon S3. You will enter your Amazon S3 connection information, directly in the Talend Data Preparation interface and create a new dataset from this data.

1. In the **Datasets** view of the Talend Data Preparation homepage, click the white arrow next to the **Add Dataset** button.
2. Select **From Amazon S3**.



The **Add an Amazon S3 dataset** form opens.

 The image shows the 'ADD AN AMAZON S3 DATASET' form. At the top, the title 'ADD AN AMAZON S3 DATASET' is displayed in a light blue header bar with a close button (X) on the right. Below the header, the form contains several fields:

- 'Dataset name*': A text input field with the placeholder text 'Amazon S3 dataset'.
- 'Specify AWS credentials': A toggle switch that is currently turned on (blue).
- 'Access key': A text input field with the placeholder text 'access key'.
- 'Secret key': A text input field with a masked placeholder consisting of dots.

 At the bottom right of the form, there is a blue button labeled 'TEST CONNECTION'.

3. In the **Dataset name** field, enter the name you want to give your dataset, Amazon S3 dataset for example.
4. Select the **Specify AWS credentials** check box.
5. Enter your Amazon S3 access key and secret key in the corresponding fields.

Amazon recommends to specify your credentials using one of the methods listed on the [Using the Default Credential Provider Chain](#) page. You will not have to manually enter your AWS credentials each time and you will be able to leave the check box unselected.

The **Amazon ECS container credentials** method from this page is not supported for Talend Data Preparation.

This procedure must be completed on the Components Catalog server, as well as the Spark Job Server if you are using Talend Data Preparation with Big Data.

6. Click **Test connection.**

If the connection is successful, the second part of the form is displayed, where you can select the object to import. If the connection is not successful, an error message is displayed, detailing why the connection failed.

The screenshot shows a form with the following fields and values:

- Region:** Please select a region (dropdown menu)
- Bucket:** MyBucket
- Object:** select/your/file
- Format*:** CSV (dropdown menu)
- Record Delimiter:** LF("\n") (dropdown menu)
- Field Delimiter:** Semicolon (dropdown menu)

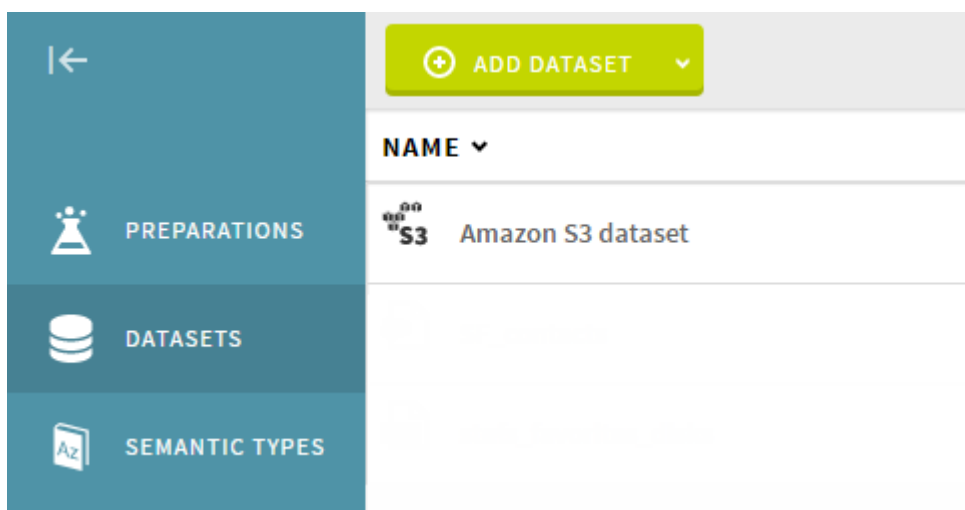
At the bottom of the form, there are two buttons: a grey **CANCEL** button on the left and a green **ADD DATASET** button on the right.

- 7.** From the **Region** and **Bucket** drop-down lists, select the location of your data in Amazon S3. You can specify a custom value for the **Region** field.
- 8.** In the **Object** field, enter the path to the dataset to import from your bucket.
- 9.** Select the format, record delimiter and field delimiter of your data in the corresponding drop-down lists.
- 10.** Click the **Add dataset** button at the end of the form.

When the import is done, the data extracted from Amazon S3 directly opens in the grid and you can start working on your preparation the same way you usually do.

The data is still stored in Amazon S3, Talend Data Preparation only retrieves a sample on-demand.

The dataset is added to the list in the **Datasets** view of the homepage.



Exporting a preparation made on an Amazon S3 dataset

When you are finished preparing a dataset extracted from Amazon S3, you may want to export your data.

1. Click the **Export** button in the application header bar.

EXPORT TO AMAZON S3



☐ Current sample ☒ All data

☐ Local CSV file

☐ Local XLSX file

☐ Local TABLEAU file

☒ Amazon S3

Access key:

access key

Secret key:

.....

Region:

Select a Region

Bucket:

MyBucket

Object:

path/to/your/object

☒ Encrypt data at rest

KMS customer master key:

KMS master key

Format:

CSV

Delimiter:

☒ Semicolon ☐ Comma ☐ Tab ☐ Space ☐ Other

Record delimiter:

☒ LF ☐ CR ☐ CRLF ☐ Other

CANCEL

CONFIRM

2. Select the **All data** checkbox.

In this example the result of the preparation is larger than the current sample size, 10000 rows by default.

3. Select **Amazon S3**.

The **Amazon S3** export is only available if the result of your preparation is larger than 10000 rows by default.

4. Enter your Amazon S3 access key and secret key in the corresponding fields.

5. Select a Region from the drop-down list and manually enter the name of the bucket where you want to store the data.

6. In the **Object** field, enter the path to the object that will store your data in the bucket.

7. If you choose to select the **Encrypt data at rest** check box to enable data encryption, enter your KMS master key.
8. Select the format and delimiters to use for the output file.
9. Click **Confirm**.

If you are using Talend Data Preparation in a Big Data context, the export will be processed on your Hadoop cluster. Else, it will be processed on the Talend Data Preparation server.

In a Big data context, preparation steps that only apply to a single row will be skipped during the export.

The export process is launched in the background. You can check the status of the export, and download your output file in the **Export history** page. For more information, see [The export history page](#) on page 150.

Export options and runtimes matrix

This table describes the export possibilities for your data and the runtime used to export your preparations according to your data source and target.

In Talend Data Preparation, two runtimes are available to process your data when you export a preparation:

- A Java runtime.
- A Big Data runtime based on Apache Beam, available when using Talend Data Preparation in a Big Data context

Depending on the data source, and the export target, the runtime used will vary, and you may or may not be able to export preparation made on large datasets. For more information, see [Working on large datasets](#) on page 83.

the different behaviors and possibilities are listed in the table below.

Input/Output	Local CSV/Excel/ Tableau file	HDFS file	Amazon S3
Local CSV/Excel/ Tableau file	Java runtime	Not available	Java runtime
Talend Job	Java runtime	Not available	Java runtime
JDBC	Java runtime	Big Data runtime	Big Data runtime if available, Java runtime otherwise
HDFS	Not available	Big Data runtime	Big Data runtime
Amazon S3	Java runtime	Big Data runtime	Big Data runtime if available, Java runtime otherwise

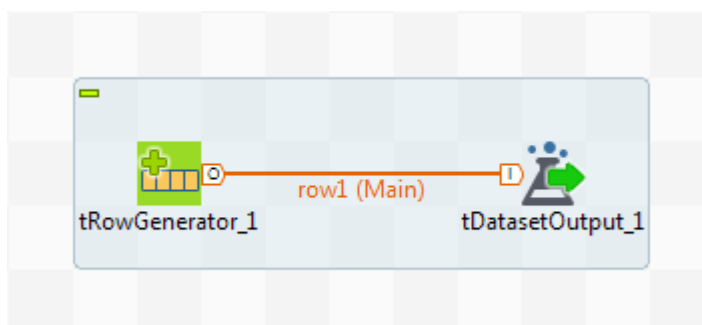
Input/Output	Local CSV/Excel/ Tableau file	HDFS file	Amazon S3
Salesforce	Java runtime	Not available	Java runtime

Developer tasks

Creating a dataset from a Talend Job

You can use a Talend Job with any input flow to create a dataset in Talend Data Preparation.

To create a dataset via Talend Studio, you must design a Job that uses the **tDatasetOutput** component as output and set it to **Create** mode. You can use any type of input flow, but the simplest Job design required to create a dataset is the following:



1. In the design workspace, add the **tRowGenerator** component, and click the **Component** tab to define its basic settings.

Schema		Functions		Preview
Column	Type	Functions	Environment varia...	Preview
id	String	TalendString.getAs...	length=>20 ;	
first_name	String	TalendString.getAs...	length=>20 ;	
last_name	String	TalendString.getAs...	length=>20 ;	

Columns Number of Rows for RowGenerator 100

2. Click the [...] next to **RowGenerator Editor** to configure a schema for your data and choose the number of rows to be generated.
3. In the design workspace, add the **tDatasetOutput** component, and click the **Component** tab to define its basic settings.

tDatasetOutput_1	
Basic settings	Schema: Built-In Edit schema Sync columns
Advanced settings	Url: "http://localhost:9999/" *
Dynamic settings	Email: "user@dataprep.com" *
View	Password: ***** *
Documentation	Mode: Create
Validation Rules	Dataset Name: "create_dataset_from_job" *
	Limit:

4. Click **Sync Column** to retrieve the schema from the previous component.
5. In the **URL** field, type the URL of the Talend Data Preparation web application, between double quotes. Port 9999 is the default port for Talend Data Preparation .
6. In the **Email** field, type the email address that you use to log in the Talend Data Preparation web application, between double quotes.
7. In the **Password** field, type your password for the Talend Data Preparation web application, between double quotes.

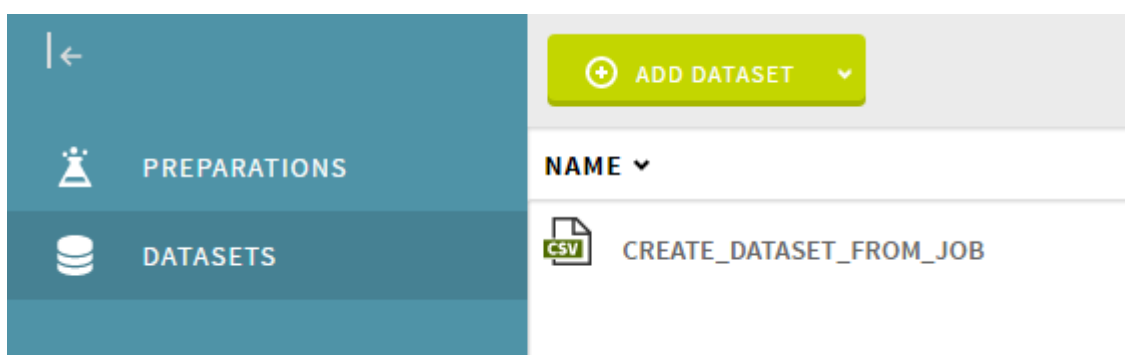
The user those credentials belong to, will be the owner of the newly created dataset. He will also be the one to have the possibility to share this dataset to other users.

8. Select the **Create** mode from the **Mode** drop-down list.

Setting the mode to **Update** allows you to use the input to update the dataset defined in the **Dataset Name** field.

9. In the **Dataset Name** field, enter a name for your dataset, between double quotes, create_dataset_from_job in this example.
10. Link the two components together using a **Row > Main** link.
11. Save your Job and press **F6** to execute it.

You can now log in the Talend Data Preparation web application, where the new dataset is available in the **Datasets** view.



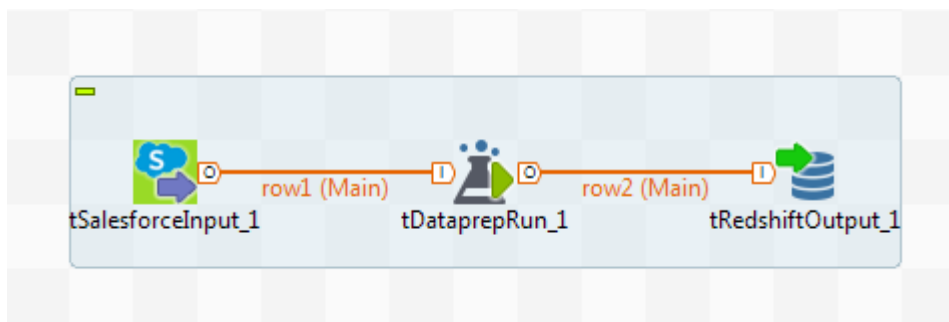
Operationalizing a recipe in a Talend Job

It is possible to use a preparation as part of a data integration flow in Talend Studio.

The **tDataprepRun** component allows you to reuse an existing preparation made in Talend Data Preparation, directly in a data integration Job. In other words, you can operationalize the process of applying a preparation to input files that have the same model.

This example shows a Job design that applies a preparation on a Salesforce input, and outputs it to a Redshift database. This assumes that a preparation has been created beforehand, on a dataset with the same schema as your input file for the Job. In this case, the existing preparation is called **datapreparun_preparation**.

The **tDataprepRun** component is an intermediary step and requires an input and an output flow. You can use any type of input and output flow, but a basic working Job design would look like the following:



In order to make the **tDataprepRun** component work when running Talend Data Preparation with an https connection, complete the following configuration:

- Retrieve Talend Data Preparation certificate, or its Certificate Authority and add it to an existing or new .jks file following this example: `keytool -import -trustcacerts -alias <cert-alias> -file <dp_certificate.crt> -keystore <truststore.jks>`
- To make the Studio trust the Talend Data Preparation certificate, edit the .ini file used to start the Studio:

```
-Djavax.net.ssl.trustStore=/path/to/<trust-store.jks>
```

```
-Djavax.net.ssl.trustStorePassword=<trust-store password>
```

- Connect a **tSetKeystore** component to **tSalesforceInput** with an **OnSubjobOk** link in order for the Job to trust the Talend Data Preparation certificate.
1. In the design workspace of Talend Studio, add a **tSalesforceInput**, a **tDataprepRun**, a **tRedshiftOutput**, and link them together using two **Row > Main** links.
 2. Select the **tSalesforceInput** component and click the **Component** tab to define its basic settings. Make sure that the schema of the **tSalesforceInput** component matches the schema expected by the **tDataprepRun** component. In other words, the input schema must be the same as the dataset upon which the preparation was made in the first place.
 3. Select the **tDataprepRun** component and click the **Component** tab to define its basic settings.

4. Enter your Talend Data Preparation connection information.
5. Click **Choose an existing preparation** to display a list of the preparations available in Talend Data Preparation.

Name	Author	Last Modification
<input checked="" type="checkbox"/> datapreun_preparation	2	28/06/16 17:23

6. Select the checkbox in front of the preparation you want to apply and click **OK**.
7. Click **Fetch Schema** to retrieve the schema of the preparation, **datapreun_preparation** in this case.

The output schema of the **tDataprepRun** component now reflects the changes made with each preparation step. The schema takes into account columns that were added or removed for example.

8. Select the **tRedshiftOutput** component and click the **Component** tab to define its basic settings.
9. Click **Sync columns** to retrieve the new output schema, inherited from the **tDataprepRun** component.
10. Save your Job and press **F6** to run it.

All the preparation steps of **datapreun_preparation** have been applied to your data, directly in the flow of your data integration Job.

Operationalizing a recipe in Talend Spark Job

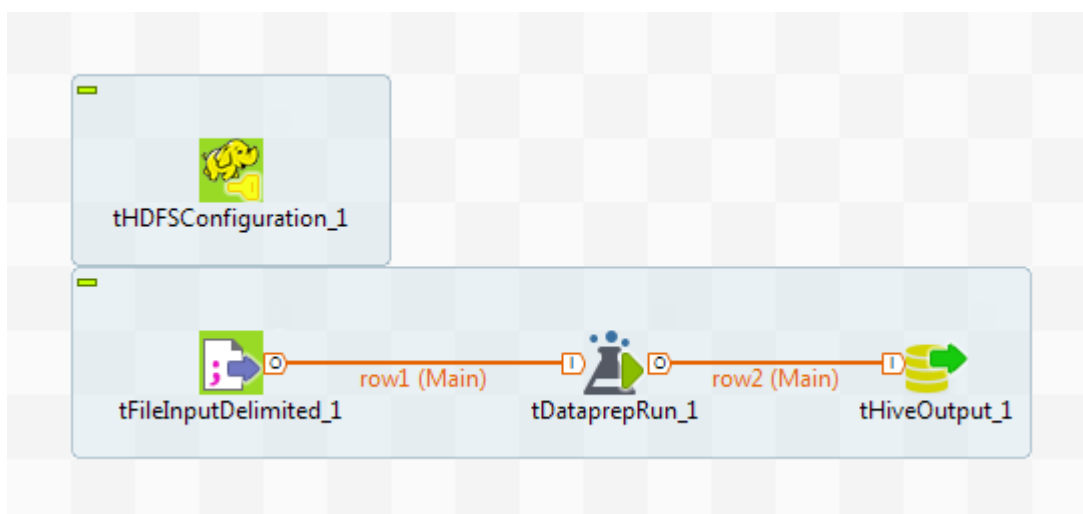
The **tDataprepRun** component allows you to reuse an existing preparation made in Talend Data Preparation, directly in a Big Data Job.

In other words, you can operationalize the process of applying a preparation to input data with the same model.

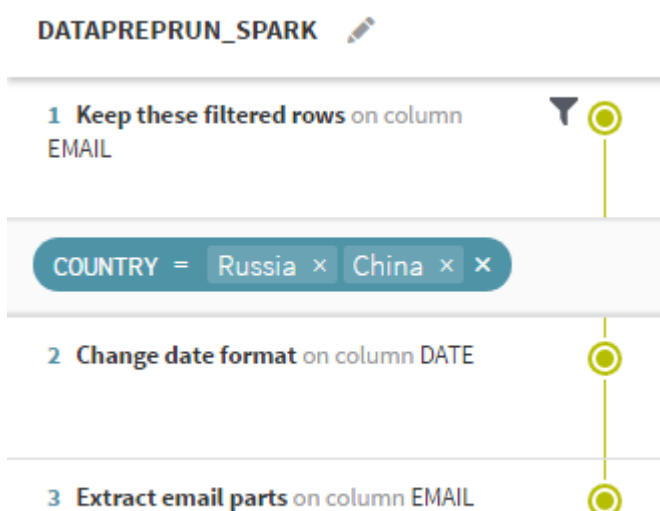
Let's take the example of a simple Job that :

- Reads customer data from a `.csv` file on HDFS,
- applies an existing preparation on this data,

- outputs it in a Hive database.



This assumes that a preparation has been created beforehand, on a dataset with the same schema as your input data for the Job. In this case, the existing preparation is called **datapreprun_spark**. This preparation was made on a dataset containing data about customers from around the world, including their names, email addresses, a subscription date, and the country they live in. This simple preparation applies a filter on the data to only keep customers from China and Russia, harmonizes the date format, and extracts the email parts.



Note that if a preparation contains actions that only affect a single row, or cells, they will be skipped by the **tDataprepRun** component during the job. The **Make as header** or **Delete Row** functions for example do not work in a Big Data context.

1. In Talend Studio, create a new Spark Batch or Spark Streaming Job.
2. In the design workspace, add a **tHDFSConfiguration**, a **tFileInputDelimited**, a **tDataprepRun** and a **tHiveOutput** component.
3. Link the **tFileInputDelimited**, **tDataprepRun** and **tHiveOutput** together using two **Row > Main** links.

4. Select the **tHDFSConfiguration** component and click the **Run** tab to configure the **Spark Configuration** tab.
5. Select the **tFileInputDelimited** component and click the **Component** tab to configure its basic settings.

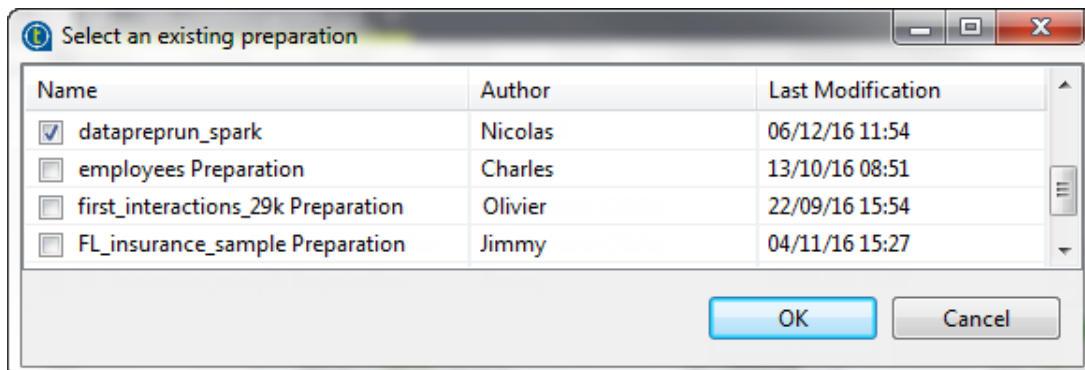
Make sure that the schema of the **tFileInputDelimited** component matches the schema expected by the **tDataprepRun** component. In other words, the input schema must be the same as the dataset upon which the `datapreprun_spark` preparation was made in the first place.

6. Select the **tDataprepRun** component and click the **Component** tab to define its basic settings.

The screenshot shows the configuration window for the **tDataprepRun** component. It is divided into two main sections: **Data Preparation Connection** and **Configuration**.

- Data Preparation Connection:**
 - URL:** `http://localhost:9999/`
 - Email:** `user@dataprep.com`
 - Password:** `*****`
- Configuration:**
 - Preparation:** `datapreprun_spark` (with a button to 'Choose an existing preparation')
 - Version:** `Current state` (with a button to 'Choose a Version')
 - Buttons: **Fetch Schema**, **Schema**, **Built-In** (dropdown), and **Edit schema** (button).

7. In the **URL** field, type the URL of the Talend Data Preparation Web application.
Port 9999 is the default port for Talend Data Preparation.
8. In the **Username** and **Password** fields, enter your Talend Data Preparation connection information, between double quotes.
9. Click **Choose an existing preparation** to display a list of the preparations available in Talend Data Preparation, and select **datapreprun_spark**.



A warning is displayed next to preparations containing incompatible actions, that only affect a single row or cell.

10. Click **Fetch Schema** to retrieve the schema of the preparation, **datapreprun_spark** in this example.

The output schema of the **tDataprepRun** component now reflects the changes made with each preparation step. The schema takes into account columns that were added or removed for example.

11. Select the **tHiveOutput** component and click the **Component** tab to define its basic settings.
12. Click **Sync columns** to retrieve the new output schema, inherited from the **tDataprepRun** component
13. Save your Job and press **F6** to run it.

All the preparation steps of `dataprepreun_spark` have been applied to your data, directly in the flow of your data integration Job.

Creating a dataset based on an on-demand Job execution

The live dataset feature allows you to create a Job in Talend Studio, execute it on demand via Talend Administration Center, and retrieve a dataset with the sample data directly in Talend Data Preparation.

Because the data originates from Talend Studio, you can take advantage of the full components palette and their Data Quality or Big Data capabilities. Unlike a local file import, where the data is stored in the Talend Data Preparation server for as long as the file exists, a live dataset only retrieves this data temporarily.

Before creating a live dataset, take the following prerequisites into account:

- It is recommended to create a remote connection for your project in Talend Studio,
- In Talend Administration Center, make sure that you are registered as a Talend Data Preparation user with the corresponding role,
- In Talend Administration Center, make sure that you have the appropriate project authorization and role.

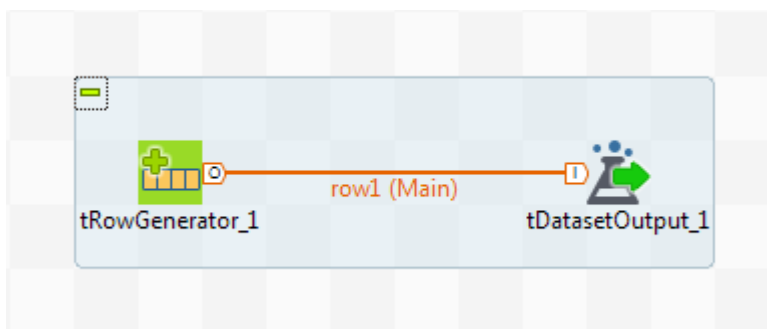
Designing the Job in Talend Studio

The first step in creating a Live dataset is to design a Job in a remote project that uses the **tDatasetOutput** component as output.

Creating a Live dataset from a `.zip` archive is also possible with a local project.

In order for dataset to be able to communicate with Talend Administration Center, it is recommended to open a project with a remote connection.

The simplest Job design required to create a working Live dataset is the following:



To create a working Live dataset when running Talend Data Preparation with an `https` connection, complete the following configuration.

- Retrieve Talend Data Preparation certificate, or its Certificate Authority and add it to an existing or new `.jks` file following this example: `keytool -import -trustcacerts -alias <cert-alias> -file <dp_certificate.crt> -keystore <truststore.jks>`
- Connect a **tSetKeystore** component to **tRowGenerator** with an **OnSubjobOk** link in order for the Job to trust the Talend Data Preparation certificate.

1. In the design workspace, add an input component, **tRowGenerator** in this example, and click the **Component** tab to define its basic settings.

Schema		Functions		Preview	
Column	Type	Functions	Environment varia...	Preview	
id	Integer	Mathematical.BIT...	a=>1 ; b=>2 ;		
first_name	String	TalendDataGenera...			
last_name	String	TalendDataGenera...			

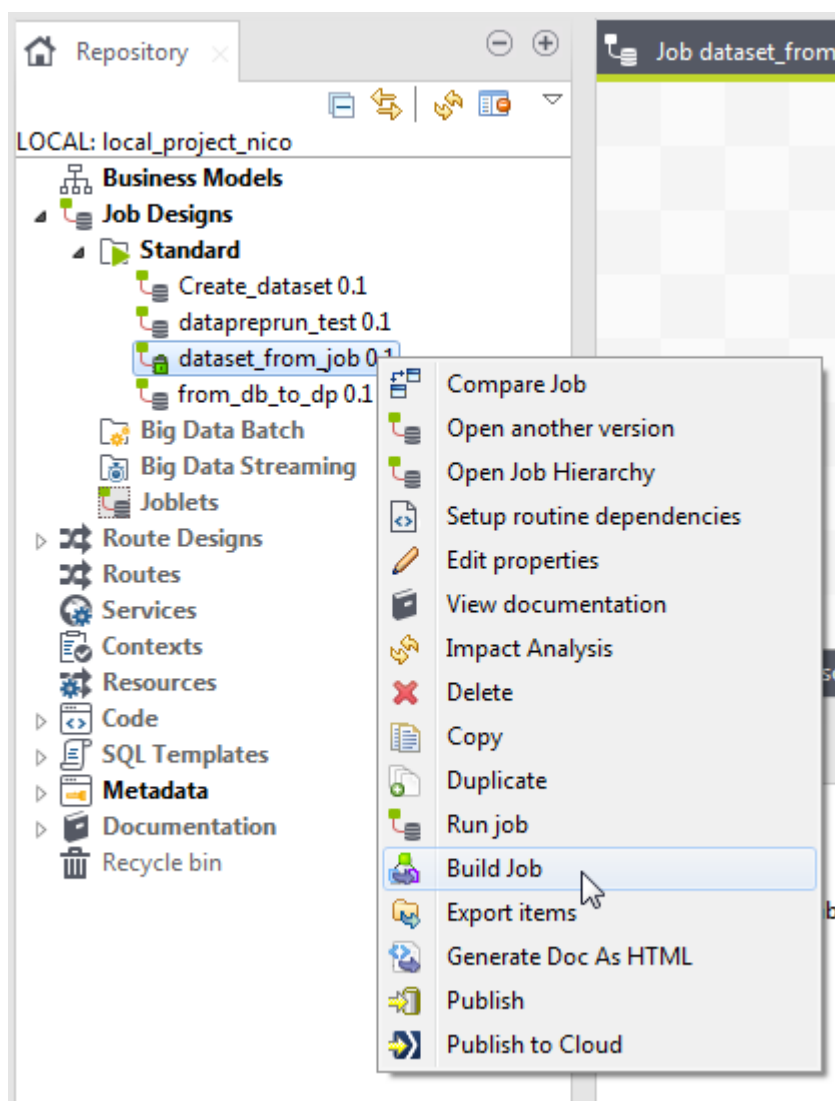
Columns ▼ Number of Rows for RowGenerator 1000

2. Click the [...] next to **RowGenerator Editor** to configure a schema for your data and choose the number of rows to be generated.
3. Add the **tDatasetOutput** component in the design workspace.
4. Link the **tRowGenerator** and **tDatasetOutput** components together using a **Row > Main** link.
5. Click the **Component** tab of the **tDatasetOutput** component to define its basic settings.

tDatasetOutput_1

Basic settings	Schema	Built-In ▼ Edit schema ... Sync columns
Advanced settings	Url	context.getProperty("dataprep_url") *
Dynamic settings	Mode	LiveDataset ▼
View	Limit	context.getProperty("dataprep_limit")
Documentation		
Validation Rules		

6. Click **Sync Column** to retrieve the schema from the previous component.
7. Select **LiveDataset** in the **Mode** list.
The **Url** and **Limit** fields are automatically filled.
8. Save your Job.
9. If you are working on a local project, right-click the name of your Job in the **Repository** tree view and click **Build Job** to export your Job as an archive that you will be able to upload in Talend Administration Center.



Creating a task in Talend Administration Center

After creating the Job that outputs your data in Talend Studio, you must import it in Talend Administration Center to create an execution task.

Before creating the execution task, you need to make sure that:

- your Talend Administration Center role is set as **Operation manager**,
- the **Data Preparation User** box is checked,
- your **Data Preparation Role** is set as **Administrator** in order to have the right to create a live dataset in Talend Data Preparation,
- you have the authorizations to read and write for the same project in which the Job was created.

Type:

Role:

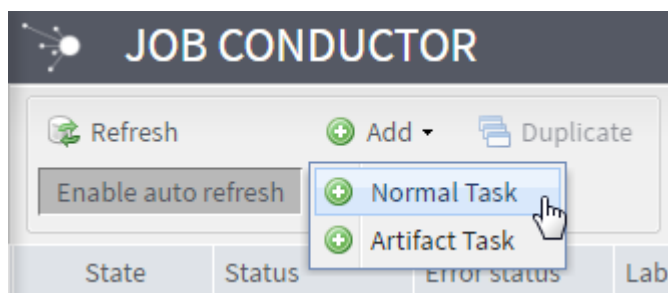
Data Preparation User: ☒

Data Preparation Role:

Group:

Active: ☒

1. Log in to Talend Administration Center.
2. In the **Job Conductor** tab, create a new execution task.

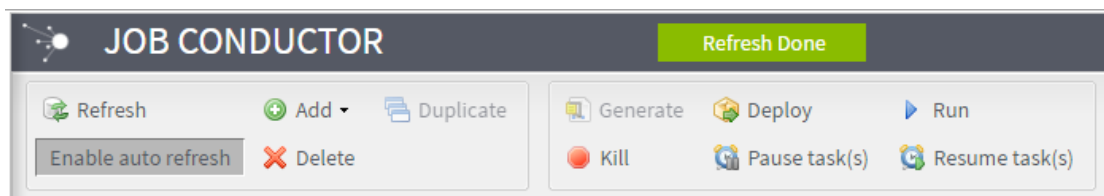


3. In the **Label** field, give a name to your execution task. The name must have dataprep_ as a prefix.

Execution task

Label:

4. In the **Job** field, import your Job either directly from Talend Studio, or via a .zip archive.
5. Select your execution server in the corresponding list and click **Save**.
6. When the status of your task is **Ready to deploy**, select it and click **Deploy**.



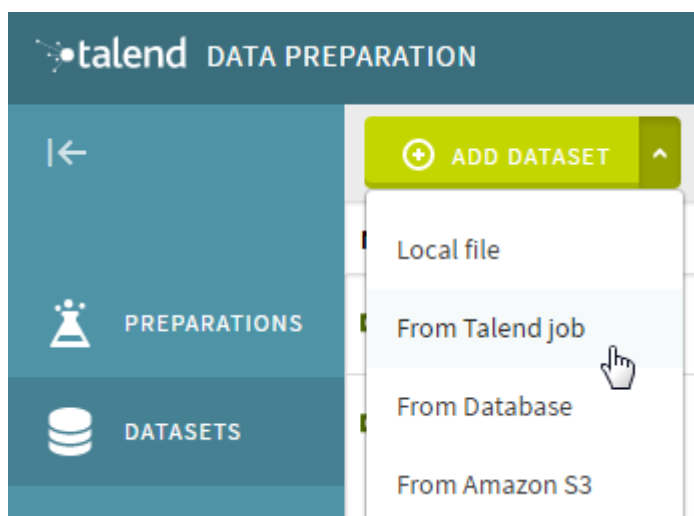
Your Talend Job can now be identified and retrieved in Talend Data Preparation.

Adding a Live Dataset in Talend Data Preparation

After creating the execution task that is ready to run in Talend Administration Center, the data can now be retrieved in Talend Data Preparation in the form of a dataset.

1. Log in to Talend Data Preparation using your user email address and password for your account.

2. Click **Datasets** to open the list of datasets.
3. Click the white arrow next to **Add Dataset** and select **From Talend Job**.



The **Add Talend Job Dataset** window opens.

4. Enter a name for the new dataset.
5. In the **User** and **Password** fields, enter your Talend Data Preparation authentication data. The credentials must belong to a user that has the right to create and run execution tasks in Talend Administration Center.
6. In the **Select the Talend Job** drop-down list, select the execution Job from which the dataset will be created.

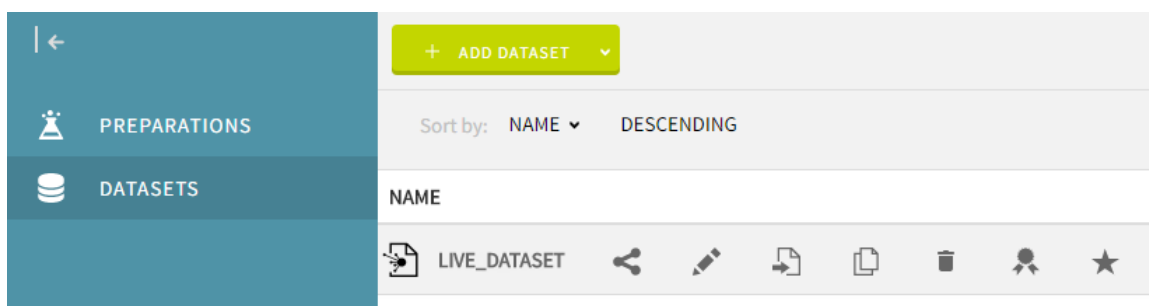
This list of Jobs correspond to the list of execution tasks that are ready to run in Talend Administration Center for the user listed in the Talend Data Preparation configuration file.

 The screenshot shows a dialog box titled 'ADD TALEND JOB DATASET' with a close button (X) in the top right corner. It contains four input fields: 'Dataset name:' with the text 'livedataset_test', 'User:' with the text 'user@dataprep.com', 'Password:' with masked characters '*****', and 'Talend job:' with a dropdown menu showing 'dataset_from_job'. At the bottom are two buttons: 'CANCEL' and 'OK'.

7. Click **OK**.

The data retrieved from the Job execution directly opens in the grid and you can start working on your preparation.

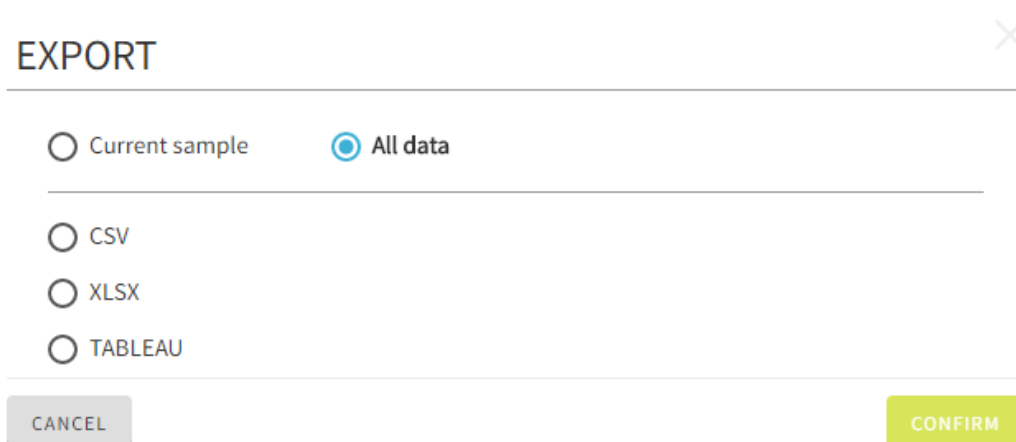
The dataset is added to the list in the **Datasets** view of the homepage.



Exporting a preparation made on a Live dataset

When you are finished preparing a dataset created from an on-demand Job execution, you may want to export your data.

1. Click the **Export** button in the application header bar.



2. Choose the file format you want to use when exporting your data.
 - If you choose **CSV**, choose a delimiter to use and enter a name for the file to export.
 - If you choose **XLSX**, or **Tableau** choose a name for the file to export.
3. Click **Confirm**.

The export operation is processed on the Talend Data Preparation server.

If your preparation was made on a dataset smaller than 10 000 rows, the download automatically starts. But if you chose to export more than 10 000 rows, you must wait for the export process to end, and download your output file in the **Export history** page. For more information, see [The export history page](#) on page 150.

The export process triggers a refresh in the data that is fetched from Job execution, guaranteeing that the data displayed in the output is always up to date.

Reference

Data Masking

Depending on the semantic type of the column on which you use the **Mask data (obfuscation)** function, the effect will vary.

The table below describes the effects of the **Mask data (obfuscation)** function on the different semantic types that support data masking.

Semantic type	Data masking effect
ADDRESS_LINE	<p>The street number is replaced by a randomly generated number and the other characters are replaced by X. However, the following key words are not transformed:</p> <p>Rue, rue, r., strasse, Strasse, Street, street, St., St, Strae, Strada, Rua, Calle, Ave., avenue, Av., Allée, allée, alle, Avenue, Avenida, Bvd., Bd., Boulevard, boulevard, Blv., Viale, Avenida, Bulevar, Route, route, road, Road, Rd., Chemin, Way, Cour, Court, Ct., Place, place, Pl., Square, Impasse, Alle, Driveway, Auahrt, Viale, Esplanade, Esplanade, Promenade, Lungomare, Esplanada, Esplanada, Faubourg, faubourg, Suburb, Vorort, Periferia, Subrbio, Suburbio, Via, Via, industrial, area, zone, industrielle, Périphérique, Peripheral, Voie, voie, Track, Gleis, Carreggiata, Caminho, Pista, Forum, STREET, RUE, ST., AVENUE, BOULEVARD, BLV., BD, ROAD, ROUTE, RD., RTE, WAY, CHEMIN, COURT, CT., SQUARE, DRIVEWAY, ALLEE, DR., ESPLANADE, SUBURB, BANLIEUE, VIA, PERIPHERAL, PERIPHERIQUE, TRACK, VOIE, FORUM, INDUSTRIAL, AREA, ZONE, INDUSTRIELLE.</p>
CITY	Replaces each character with a random one.
COMPANY	Generates a random but existing company name.
DECIMAL	Replaces each number with a random one.
EMAIL	Replaces everything before the @ character with X, and leaves the rest untransformed.
FIRST_NAME	Generates a random first name.
LAST_NAME	Generates a random last name.
FULL_NAME	Generates a random first name and last name.
FR_COMMUNE	Generates a random french city name

Semantic type	Data masking effect
INTEGER	Replaces each number with a random one.
IPv4_ADDRESS	Generates a correct random IPv4 address
IPv6_ADDRESS	Generates a correct random IPv6 address
JOB_TITLE	Replaces each character with a random one.
LOCALIZATION	Generates random longitude and latitude coordinates.
LOCATION_COORDINATE	Generates random longitude and latitude coordinates.
MAC_ADDRESS	Generates a correct random MAC address.
ORGANIZATION	Generates a correct random organization name.
PASSPORT	Generates a correct random passport number.
US_PHONE	Generates a correct random phone number for the US.
FR_PHONE	Generates a correct random phone number for France.
UK_PHONE	Generates a correct random phone number for the UK.
DE_PHONE	Generates a correct random phone number for Germany.
US_POSTAL_CODE	Generates a correct random postal code for the US.
FR_POSTAL_CODE	Generates a correct random postal code for France.
UK_POSTAL_CODE	Generates a correct random postal code for the UK.
DE_POSTAL_CODE	Generates a correct random postal code for Germany.
BE_POSTAL_CODE	Generates a correct random postal code for Belgium.
FR_CODE_COMMUNE_INSEE	Generates a random french INSEE city code.
US_SSN	Generates a correct random Social Security Number for the US.

Semantic type	Data masking effect
FR_SSN	Generates a correct random Social Security Number for France.
UK_SSN	Generates a correct random Social Security Number for the UK.
TEXT	Replaces each character with a random one.
MASTERCARD	Generates a correct random MasterCard credit card number.
US_CREDIT_CARD	Generates a correct random American Express credit card number.
VISACARD	Generates a correct random Visa credit card number.

Date Formats

Depending on the countries, dates may have different formats. Here are the elements that can help you change the date format or customize it.

Date format	Description
EEEE	Day of the week
MM	Month of the year in a two-digit format
MMM	Abbreviated month of the year
MMMM	Month of the year
dd	Day of the month
yy	Year in two-digit format
yyyy	Year in four-digit format
YYYY	Week-based year
HH	Hour of the day (0-23)
hh	Clock hour in AM/PM (1-12) format
mm	Minute
ss	Second
SSS	Fraction of a second
a	AM/PM marker

Date format	Description
Z	UTC time offset

List of date and date/time formats

Date and time are formatted according to the different conventions all around the world. Locale-specific date and date/time formats are specified by date and time pattern strings. The following tables provide information on the patterns which are recognized as date or date/time data in Talend Data Preparation.

According to the locale of your Java installation, the validation results may be different when using patterns which have a weekday and a month name. Talend Data Preparation first validates dates and times using the `en_US` locale and then, using the default locale of the Java installation of the server.

For example, if the Java Virtual Machine locale is set to French, "22. March 1999" and "22. mars 1999" will be valid dates but not "22. März 1999", even though `d. MMMM yyyy` is originally a German pattern.

ISO 8601 patterns

Date and time patterns	Example
yyyyMMddZ	19990322+0100
yyyyMMdd	19990322
yyyy-MM-dd G	1999-03-22 AD
yyyy-MM-ddXXX	1999-03-22+01:00
yyyy-MM-dd'T'HH:mm:ss.SSS['VV']	1999-03-22T05:06:07.000[Europe/Paris]
yyyy-MM-dd'T'HH:mm:ss.SSS	1999-03-22T05:06:07.000
yyyy-MM-dd'T'HH:mm:ss	1999-03-22T05:06:07
yyyy-MM-dd'T'HH:mm:ss.SSS'Z'	1999-03-22T05:06:07.000Z
yyyy-MM-dd'T'HH:mm:ss.SSSXXX	1999-03-22T05:06:07.000+01:00
yyyy-MM-dd'T'HH:mm:ssXXX	1999-03-22T05:06:07+01:00
yyyy-DDDXXX	1999-081+01:00
YYYY'W'wc	1999W132

Date and time patterns	Example
YYYY-'W'w-c	1999-W13-2
yyyy-MM-dd'T'HH:mm:ss.SSSXXX['VV']	1999-03-22T05:06:07.000+01:00[Europe/Paris]
yyyy-MM-dd'T'HH:mm:ssXXX['VV']	1999-03-22T05:06:07+01:00[Europe/Paris]

Locale be: Belarussian

Date and time patterns	Example
d.M.yy	22.3.99
d.M.yy H.mm	22.3.99 5.06
d.M.yyyy H.mm.ss	22.3.1999 5.06.07

Locale cs: Czech

Date and time patterns	Example
d.M.yyyy H:mm:ss	22.3.1999 5:06:07

Locale da: Danish

Date and time patterns	Example
dd-MM-yy	22-03-99
dd-MM-yy HH:mm	22-03-99 05:06
dd-MM-yyyy HH:mm:ss	22-03-1999 05:06:07

Locale de_DE: German, Germany

Date and time patterns	Example
dd.MM.yy	22.03.99

Date and time patterns	Example
d. MMMM yyyy	22. März 1999
EEEE, d. MMMM yyyy	Montag, 22. März 1999
dd.MM.yyyy	22.03.1999
dd.MM.yy HH:mm	22.03.99 05:06
d. MMMM yyyy HH:mm:ss z	22. März 1999 05:06:07 MEZ
dd.MM.yyyy HH:mm:ss	22.03.1999 05:06:07
dd.MM.yy HH:mm:ss	22.03.99 05:06:07
dd.MM.yyyy HH:mm	22.03.1999 05:06
EEEE, d. MMMM yyyy HH:mm' Uhr 'z	Montag, 22. März 1999 05:06 Uhr MEZ

Locale en_CA: English, Canada

Date and time patterns	Example
d-MMM-yyyy	22-Mar-1999
dd/MM/yy h:mm a	22/03/99 5:06 AM
EEEE, MMMM d, yyyy h:mm:ss 'o'clock' a z	Monday, March 22, 1999 5:06:07 o'clock AM CET
d-MMM-yyyy h:mm:ss a	22-Mar-1999 5:06:07 AM

Locale en_GB: English, United Kingdom

Date and time patterns	Example
dd MMMM yyyy	22 March 1999
EEEE, d MMMM yyyy	Monday, 22 March 1999

Date and time patterns	Example
dd-MMM-yyyy	22-Mar-1999
dd MMMM yyyy HH:mm:ss z	22 March 1999 05:06:07 CET
EEEE, d MMMM yyyy HH:mm:ss 'o'clock' z	Monday, 22 March 1999 05:06:07 o'clock CET
dd-MMM-yyyy HH:mm:ss	22-Mar-1999 05:06:07
dd-MMM-yy hh.mm.ss.nnnnnnnnn a	22-Mar-99 05.06.07.000000888 AM

Locale en_US : English, United States

Date and time patterns	Example
M/d/yy	3/22/99
MM/dd/yy	03/11/22
MM-dd-yy	03-22-99
M-d-yy	3-22-99
MMM d, yyyy	Mar 22, 1999
MMMM d, yyyy	March 22, 1999
EEEE, MMMM d, yyyy	Monday, March 22, 1999
MMM d yyyy	Mar 22 1999
MMMM d yyyy	March 22 1999
MM-dd-yyyy	03-22-1999
M-d-yyyy	3-22-1999
yyyy-MM-ddXXX	1999-03-22+01:00
dd/MM/yyyy	22/03/1999

Date and time patterns	Example
d/M/yyyy	22/3/1999
MM/dd/yyyy	03/22/1999
M/d/yyyy	3/22/1999
yyyy/M/d	1999/3/22
M/d/yy h:mm a	3/22/99 5:06 AM
MM/dd/yy h:mm a	03/22/99 5:06 AM
MM-dd-yy h:mm a	03-22-99 5:06 AM
M-d-yy h:mm a	3-22-99 5:06 AM
MMM d, yyyy h:mm:ss a	Mar 22, 1999 5:06:07 AM
EEEE, MMMM d, yyyy h:mm:ss a z	Monday, March 22, 1999 5:06:07 AM CET
EEE MMM dd HH:mm:ss z yyyy	Mon Mar 22 05:06:07 CET 1999
EEE, d MMM yyyy HH:mm:ss Z	Mon, 22 Mar 1999 05:06:07 +0100
d MMM yyyy HH:mm:ss Z	22 Mar 1999 05:06:07 +0100
MM-dd-yyyy h:mm:ss a	03-22-1999 5:06:07 AM
M-d-yyyy h:mm:ss a	3-22-1999 5:06:07 AM
yyyy-MM-dd h:mm:ss a	1999-03-22 5:06:07 AM
yyyy-M-d h:mm:ss a	1999-3-22 5:06:07 AM
yyyy-MM-dd HH:mm:ss.S	1999-03-22 05:06:07.0
dd/MM/yyyy h:mm:ss a	22/03/1999 5:06:07 AM
d/M/yyyy h:mm:ss a	22/3/1999 5:06:07 AM
MM/dd/yyyy h:mm:ss a	03/22/1999 5:06:07 AM

Date and time patterns	Example
M/d/yyyy h:mm:ss a	3/22/1999 5:06:07 AM
MM/dd/yy h:mm:ss a	03/22/99 5:06:07 AM
MM/dd/yy H:mm:ss	03/22/99 5:06:07
M/d/yy H:mm:ss	3/22/99 5:06:07
dd/MM/yyyy h:mm a	22/03/1999 5:06 AM
d/M/yyyy h:mm a	22/3/1999 5:06 AM
MM/dd/yyyy h:mm a	03/22/1999 5:06 AM
M/d/yyyy h:mm a	3/22/1999 5:06 AM
MM-dd-yy h:mm:ss a	03-22-99 5:06:07 AM
M-d-yy h:mm:ss a	3-22-99 5:06:07 AM
MM-dd-yyyy h:mm a	03-22-1999 5:06 AM
M-d-yyyy h:mm a	3-22-1999 5:06 AM
yyyy-MM-dd h:mm a	1999-03-22 5:06 AM
yyyy-M-d h:mm a	1999-3-22 5:06 AM
MMM.dd.yyyy	Mar.22.1999
d/MMM/yyyy H:mm:ss Z	22/Mar/1999 5:06:07 +0100
dd/MMM/yy h:mm a	22/Mar/99 5:06 AM

Locale es: Spanish

Date and time patterns	Example
d/MM/yy	22/03/99

Date and time patterns	Example
d/MM/yy H:mm	22/03/99 5:06
d.M.yy H:mm	22.3.99 5:06

Locale et: Estonian

Date and time patterns	Example
d.MM.yy	22.03.99
d.MM.yyyy	22.03.1999
d.MM.yy H:mm	22.03.99 5:06
d.MM.yyyy H:mm:ss	22.03.1999 5:06:07

Locale fi: Finnish

Date and time patterns	Example
d.M.yyyy	22.3.1999
d.M.yyyy H:mm	22.3.1999 5:06

Locale fr_CA: French, Canada

Date and time patterns	Example
yy-MM-dd	99-03-22
yyyy-MM-dd	1999-03-22
d MMMM yyyy HH:mm:ss z	22 mars 1999 05:06:07 CET
MMMM d, yyyy h:mm:ss z a	March 22, 1999 5:06:07 CET AM
yyyy-MM-dd HH:mm:ss	1999-03-22 05:06:07

Date and time patterns	Example
EEEE d MMMM yyyy H' h 'mm z	lundi 22 mars 1999 5 h 06 CET

Locale fr_FR: French, France

Date and time patterns	Example
dd/MM/yy	22/03/99
d MMM yyyy	22 mars 1999
d MMMM yyyy	22 mars 1999
EEEE d MMMM yyyy	lundi 22 mars 1999
dd/MM/yy HH:mm	22/03/99 05:06
MM/dd/yy HH:mm	03/22/99 05:06
M/d/yy HH:mm	3/22/99 05:06
MM-dd-yy HH:mm	03-22-99 05:06
M-d-yy HH:mm	3-22-99 05:06
d MMM yyyy HH:mm:ss	22 mars 1999 05:06:07
d MMMM yyyy HH:mm:ss z	22 mars 1999 05:06:07 CET
MM-dd-yyyy HH:mm:ss	03-22-1999 05:06:07
M-d-yyyy HH:mm:ss	3-22-1999 05:06:07
yyyy-M-d HH:mm:ss	1999-3-22 05:06:07
dd/MM/yyyy HH:mm:ss	22/03/1999 05:06:07
d/M/yyyy HH:mm:ss	22/3/1999 05:06:07
MM/dd/yyyy HH:mm:ss	03/22/1999 05:06:07

Date and time patterns	Example
M/d/yyyy HH:mm:ss	3/22/1999 05:06:07
EEEE d MMMM yyyy HH' h 'mm z	lundi 22 mars 1999 05 h 06 CET
dd/MM/yy HH:mm:ss	22/03/99 05:06:07
MM/dd/yy HH:mm:ss	03/22/99 05:06:07
M/d/yy HH:mm:ss	3/22/99 05:06:07
dd/MM/yyyy HH:mm	22/03/1999 05:06
d/M/yyyy HH:mm	22/3/1999 05:06
MM/dd/yyyy HH:mm	03/22/1999 05:06
M/d/yyyy HH:mm	3/22/1999 05:06
MM-dd-yy HH:mm:ss	03-22-99 05:06:07
M-d-yy HH:mm:ss	3-22-99 05:06:07
MM-dd-yyyy HH:mm	03-22-1999 05:06
M-d-yyyy HH:mm	3-22-1999 05:06
yyyy-M-d HH:mm	1999-3-22 05:06

Locale ga: Irish

Date and time patterns	Example
yy/MM/dd HH:mm	99/03/22 05:06

Locale hr: Croatian

Date and time patterns	Example
yyyy.MM.dd	1999.03.22

Date and time patterns	Example
yyyy.MM.dd HH:mm:ss	1999.03.22 05:06:07
yyyy.MM.dd HH:mm	1999.03.22 05:06

Locale hu: Hungarian

Date and time patterns	Example
yyyy.MM.dd.	1999.03.22.
yyyy.MM.dd. H:mm:ss	1999.03.22. 5:06:07
yyyy.MM.dd. H:mm	1999.03.22. 5:06

Locale is: Icelandic

Date and time patterns	Example
d.M.yyyy HH:mm:ss	22.3.1999 05:06:07
d.M.yyyy HH:mm	22.3.1999 05:06

Locale it_IT: Italian, Italy

Date and time patterns	Example
d-MMM-yyyy	22-mar-1999
dd/MM/yy H.mm	22/03/99 5.06
yy-MM-dd HH:mm	99-03-22 05:06
d-MMM-yyyy H.mm.ss	22-mar-1999 5.06.07
d MMMM yyyy H.mm.ss z	22 marzo 1999 5.06.07 CET
EEEE d MMMM yyyy H.mm.ss z	lunedì 22 marzo 1999 5.06.07 CET

Locale iw: Hebrew

Date and time patterns	Example
HH:mm dd/MM/yy	05:06 22/03/99
HH:mm:ss dd/MM/yyyy	05:06:07 22/03/1999

Locale ja_JP: Japanese, Japan

Date and time patterns	Example
yy/MM/dd	99/03/22
yyyy/MM/dd	1999/03/22
yy/MM/dd H:mm	99/03/22 5:06
MM/dd/yy H:mm	03/22/99 5:06
M/d/yy H:mm	3/22/99 5:06
MM-dd-yy H:mm	03-22-99 5:06
M-d-yy H:mm	3-22-99 5:06 AM
MM-dd-yyyy H:mm:ss	03-22-1999 5:06:07
M-d-yyyy H:mm:ss	3-22-1999 5:06:07
yyyy-MM-dd H:mm:ss	1999-03-22 5:06:07
yy/MM/dd H:mm:ss	99/03/22 5:06:07
M/d/yy h:mm:ss a	3/22/99 5:06:07 AM
yyyy/MM/dd H:mm	1999/03/22 5:06
dd/MM/yyyy H:mm	22/03/1999 5:06
d/M/yyyy H:mm	22/3/1999 5:06

Date and time patterns	Example
MM/dd/yyyy H:mm	03/22/1999 5:06
M/d/yyyy H:mm	3/22/1999 5:06
MM-dd-yy H:mm:ss	03-22-99 5:06:07
M-d-yy H:mm:ss	3-22-99 5:06:07
MM-dd-yyyy H:mm	03-22-1999 5:06
M-d-yyyy H:mm	3-22-1999 5:06
yyyy-MM-dd H:mm	1999-03-22 5:06
yyyy-M-d H:mm	1999-3-22 5:06

Locale ko: Korean

Date and time patterns	Example
yy. M. d	99. 3. 22
yyyy. M. d	1999. 3. 22

Locale lt: Lithuanian

Date and time patterns	Example
yy.M.d	99.3.22
yy.M.d HH.mm	99.3.22 05.06
yyyy-MM-dd HH.mm.ss	1999-03-22 05.06.07

Locale lv: Latvian

Date and time patterns	Example
yy.d.M	99.22.3
yyyy.d.M	1999.22.3
yy.d.M HH:mm	99.22.3 05:06
yyyy.d.M HH:mm:ss	1999.22.3 05:06:07

Locale mk: Macedonian

Date and time patterns	Example
d.M.yy HH:mm	22.3.99 05:06
d.M.yyyy HH:mm:	22.3.1999 05:06:

Locale nl: Dutch

Date and time patterns	Example
d-M-yy	22-3-99
d-M-yy H:mm	22-3-99 5:06

Locale pt: Portuguese

Date and time patterns	Example
dd-MM-yyyy	22-03-1999
dd-MM-yyyy H:mm	22-03-1999 5:06

Locale ru: Russian

Date and time patterns	Example
dd.MM.yy H:mm	22.03.99 5:06
dd.MM.yyyy H:mm:ss	22.03.1999 5:06:07

Locale sq: Albanian

Date and time patterns	Example
yyyy-MM-dd h:mm:ss.a	1999-03-22 5:06:07.PD
yy-MM-dd h.mm.a	99-03-22 5.06.PD
yyyy-MM-dd h.mm.ss.a z	1999-03-22 5.06.07.PD CET

Locale sr: Serbian

Date and time patterns	Example
d.M.yy.	22.3.99.
dd.MM.yyyy.	22.03.1999.
d.M.yy. HH.mm	22.3.99. 05.06
dd.MM.yyyy. HH.mm.ss	22.03.1999. 05.06.07
dd.MM.yyyy. HH.mm.ss z	22.03.1999. 05.06.07 CET

Locale vi: Vietnamese

Date and time patterns	Example
HH:mm:ss dd-MM-yyyy	05:06:07 22-03-1999
HH:mm dd/MM/yyyy	05:06 22/03/1999

Locale zh_CN: Chinese (Simplified), China

Date and time patterns	Example
yy-M-d	99-3-22
yyyy-M-d	1999-3-22
yyyy'#M'#d'#'	1999#3#22#
yyyy'#M'#d'#' EEEE	1999#3#22# ###
yy-M-d ah:mm	99-3-22 ##5:06
yyyy-M-d H:mm:ss	1999-3-22 5:06:07
yyyy'#M'#d'#' ahh'#mm'#ss'#'	1999#3#22# ##05#06#07#
yyyy'#M'#d'#' H'#mm'#ss'#' z	1999#3#22# 5#06#07# CET
yyyy'#M'#d'#' EEEE ahh'#mm'#ss'#' z	1999#3#22# ### ##05#06#07# CET

List of functions

This table lists all the functions available in Talend Data Preparation and their effects.

Name	Category	Description
Negate value	boolean	Reverse the boolean value of cells from this column
Change data type	column metadata	Change type of this column (number, text, date, etc.)
Change semantic domain	column metadata	Change semantic domain of this column (city, zipcode, last name, etc.)
Create new column	column metadata	Copy a column or create a brand new one
Delete column	column metadata	Delete the selected columns
Duplicate column	column metadata	Create an exact copy of this column
Rename column	column metadata	Rename this column

Name	Category	Description
Concatenate with	columns	Merge the content of this column with another one, and displays it in a new column
Reorder columns	columns	Change column order
Swap columns	columns	Swap the values with an other column
Convert distance	conversions	Convert distance from one unit to another
Convert duration	conversions	Convert duration from one unit to another
Convert temperature	conversions	Convert temperature measurement units
Clear on matching value	data cleansing	Clear cells that match the value
Clear the cells with invalid values	data cleansing	Clear cells that contain a value recognized as invalid
Delete row	data cleansing	Delete this line
Delete the rows that match	data cleansing	Delete rows where a cell in this column has a specific value
Delete the rows with empty cell	data cleansing	Delete rows that have empty cells
Delete the rows with invalid cell	data cleansing	Delete rows which contain an invalid cell
Fill cells with value	data cleansing	Fill cells from this column with a given value
Fill empty cells with text	data cleansing	Fill empty cells from this column with a given value
Fill empty cells with value	data cleansing	Fill cells from this column with a given value
Make as header	data cleansing	Cells of this line will become columns names, the line will be deleted
Remove negative values	data cleansing	Lines with a negative value in this column will be deleted

Name	Category	Description
Mask data (obfuscation)	data masking	Mask data according to the domain information of the column (anonymisation)
Lookup	data blending	Blends columns from another dataset into this one
Calculate time since	dates	Calculate elapsed time since a date in the desired unit (year, month, day, hour)
Calculate timestamp to date	dates	Given a timestamp (elapsed time since epoch in second), create a new column with the date
Change date format	dates	Change the date format to use in a date column
Compare dates	dates	Compare this column to another column or a constant
Convert dates	dates	Convert dates from one calendar to another
Extract date parts	dates	Create columns with year, month, day, hour, minute, second, etc.
Modify Dates	dates	Add or subtract time unit amount
Delete these filtered lines	filtered	Delete only the lines that match the current filters
Keep these filtered lines	filtered	Keep only the lines that match the current filters
Add, multiply, subtract or divide	math	Perform an operation/calculation on this column with another one or with a fixed value: Add/sum (+), multiply (x), subtract (-), or divide(/)
Base 10 Logarithm	math	Compute the base 10 logarithm from a column
Calculate absolute value	math	Calculate the absolute value for all the numeric values in this column.

Name	Category	Description
Cosine	math	Compute the trigonometric Cosine from a column
Exponential	math	Exponential of a column number
Max	math	Max with another column or a constant
Min	math	Min with another column or a constant
Natural logarithm	math	Compute the natural logarithm from a column
Negate	math	Negate a column number
Power	math	Power with another column or a constant
Sine	math	Compute the trigonometric Sine from a column
Square root	math	Square root of a column number
Tangent	math	Compute the trigonometric Tangent from a column
Compare numbers	numbers	Compare this column to another column or a constant
Format numbers	numbers	Allow to format number (decimal, integer & scientific) in a specific format or pattern
Remove fractional part	numbers	Round towards zero. (3.74 -> 3) and (-3.74 -> -3)
Round value using ceil mode	numbers	Round up value to the nearest number, depending on the precision you set. (3.14 -> 4 if Precision is set to 0, and 3.14 -> 3.2 if Precision is set to 1)
Round value using down mode	numbers	Round towards zero. (3.74 -> 3 and -3.74 -> -3 for a Precision set to 0)
Round value using floor mode	numbers	Round down value to the nearest number, depending on the precision you set. (3.74 -> 3 if Precision is set to 0, and 3.74 -> 3.7 if Precision is set to 1)

Name	Category	Description
Round value using halfUp mode	numbers	Round value to the closest number, depending on the precision you set. (3.14 -> 3 and 3.74 -> 4 for a Precision set to 0)
Format phone number	phones	Format a phone number to standard formats
Extract email Parts	split	Extract local and domain parts from an email
Extract number	split	Extract number from the input
Extract string parts	split	Extract string tokens based on regex groups
Extract URL Parts	split	Extract protocol, host, port, query, etc... from an URL in separated columns
Split the text in parts	split	Split column from separators
Calculate length	strings	Extract the number of digits from a value (23562 -> 5)
Change to lower case	strings	Converts all of the cell text in this column to lower case
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)
Contains text	strings	Checks if the cell contains the specified value
Extract parts of the text	strings	Extract some parts of the text (substring) and create a new column
Match similar text	strings	Create a new column with <i><i>true</i></i> or <i><i>>false</i></i> regarding if the value is less or equals the Levenshtein distance of a given value
Matches pattern	strings	Create a new column with <i><i>true</i></i> or <i><i>>false</i></i>

Name	Category	Description
		regarding if the value that matches or not a given pattern
Remove consecutive characters	strings	Remove consecutive repeated characters
Remove part of the text	strings	Remove specified text from cells in this column
Remove trailing and leading characters	strings	Remove trailing and leading spaces or other specified characters (i.e. trim)
Replace the cells that match	strings	Replace the cells that have a specific value
Add extra characters	strings advanced	Add extra characters (padding) on the left or on the right of the original value to match an expected size
Find and group similar text	strings advanced	Replace all similar values with the right one (i.e. cluster on fuzzy matching)
Remove all non alpha numeric characters	strings advanced	For example <code>€f&çÃ</code> <code>€\$Ã,Ã-10.5k</code> will become 105
Remove all non numeric characters	strings advanced	For example <code>€f&çÃ</code> <code>€\$Ã,Ã-10.5k</code> will become 10.5
Simplify text (remove case, accent, etc.)	strings advanced	Simplify the content of this column (ie: François -> francois)

Number Formats

Depending on the countries, numbers may have different formats. Here are the different formats used in Talend Data Preparation.

Format	Description	Thousands separator	Decimal separator
1,545.12 (US, UK, China, etc...)	This standard is mainly used in English-speaking countries and other countries such as China.	comma	period
1 545,12 (France, Russia, etc...)	This standard is mainly used in non English-speaking countries.	space	comma

Format	Description	Thousands separator	Decimal separator
1'545,12 (Swiss)	This standard is used in Switzerland.	apostrophe	comma
Scientific	This standard is used to write numbers using scientific notation. In this standard, 3e2 represents three times ten raised to the power of two, or 300.	comma	period

Operators Reference

When creating a pattern to be matched, several operators can be used.

Operator	Description
Equals	The cells exactly corresponding to the text you entered are matched.
Contains	The cells containing the text you entered are matched.
Starts with	The cells which contents starts with the text you entered are matched.
Ends with	The cells which contents ends with the text you entered are matched.
RegEx	The cells corresponding the regular expression you defined are matched. For a regular expression cheat sheet, see Regular Expressions Cheat Sheet .

Regular Expressions

Regular expressions (or regex) are advanced search strings that allows you to match complex patterns.

In this documentation, the regular expression elements are classified by category.

All the examples listed are used with the two following lines:

Comment from happy_user@company.com (04-Apr-2016):

I love working with Talend Data Preparation! It really helps me with all my daily tasks!

Regular Expressions Examples

Regular Expression	Matches
\bTa	Talend

Regular Expression	Matches
<code>\bw\w*</code>	working, with
<code>\w+n\b</code>	Preparation
<code>Talend\s\w+\s\w+</code>	Talend Data Preparation
<code>task(s?)</code>	tasks (it would also match "task")
<code>\w+@\w+\.com</code>	happy_user@company.com
<code>\d{2}-.*-\d+</code>	04-Apr-2016

Anchors

Character	Matches	Example
<code>^</code>	Start of string, or start of line in a multi-line pattern	<code>^Comment</code> matches "Comment" at the beginning of the line. <code>^C.*</code> matches the first line.
<code>\$</code>	End of string, or end of line in a multi-line pattern	<code>!\$</code> matches the last exclamation mark.
<code>\b</code>	Word boundary	<code>\bwo</code> matches the "wo" in "working". <code>\bwo\w+</code> matches "working". <code>ng\b</code> matches the "ng" in "working". <code>\w+ng\b</code> matches "working".
<code>\B</code>	Not word boundary	<code>\Bh</code> matches the final "h" in "with" but not the "h" in "helps" or "happy".

Character	Matches	Example
		h\B matches the first "h" in "helps" and "happy" but not the final one in "with".

Character Classes

Character	Matches	Example
.	Any character, except new line (\n)	. matches all the characters in the text, except for the carriage return.
\s	White space	Talend\sData matches "Talend Data". Data\s+Preparation matches "Data Preparation".
\S	Not white space	\S matches all the characters in the sentence, except for the spaces.
\d	Digit	\d{4} matches "2016".
\D	Not digit	\D matches all the characters in the text but not the numbers.
\w	Word character and underscore	T\w+ matches "Talend".
\W	Not word	company\Wcom matches "company.com".
\n	New line	. *\n . * matches the whole text.

Escape Characters

Character	Matches
\.	.
\\	\
\+	+
*	*
\?	?
\\$	\$
\[[
\]]
\{	{
\}	}
\((
\))
\	
\/	/

Groups and Ranges

Character	Matches	Example
()	Group	m(e y) matches "me" and "my".
(a b)	a or b	m(e y) matches "me" (in "Comment"), "me" and "my".

Character	Matches	Example
[abc]	Range (a or b or c)	m[ey] matches "me" (in "Comment"), "me" and "my".
[a-q]	Letter from a to q	m[a-m] matches "me" (in "Comment") and "me" but not "my".
[0-7]	Digit from 0 to 7	201[0-5] does not match "2016" but would match all years between "2010" and "2015".

The expression captured in a group can be reused using the \$ symbol. When more than one group is captured, add a number to the \$ symbol, so that it corresponds to the order in which they were captured.

For example, you want to reformulate the expression Y16Q02 that can be matched by the regular expression `Y(\d{2})Q(\d{2})`. You can then reformulate your original expression only keeping the characters you have captured. If you want your new expression to be `Quarter 02 of year 2016`, the new regular expression `Quarter $2 of year 20$1` will match it.

Quantifiers

Character	Matches	Examples
*	0 or more	work\w* matches "working" but also "work" and "works".
+	1 or more	work\w+ matches "working" but also "works". However, it does not match "work".
?	0 or 1	work(s?) matches "work" and "works" but not "working".

Character	Matches	Examples
{ 3 }	Exactly 3	20\d{ 2 } matches "2016" and other numbers between "2000" and "2099".
{ 3 , }	3 or more	20\d{ 2 , } matches "2016" and all numbers superiors or equal to "2000" starting by "20".
{ 3 , 5 }	3, 4 or 5	20{ 1 , 2 } matches "2016" and all numbers from "200" to "2099".
[0 - 7]	Digit from 0 to 7	201[0 - 9] matches "2016" and all numbers from "2010" to "2019".

Predefined Semantic Types

When you add a dataset, Talend Data Preparation automatically suggests one of the supported semantic types for each column.

The table below lists all the supported semantic types.

Semantic type	Description
ADDRESS_LINE	Street number and name
AIRPORT	Airport
AIRPORT_CODE	Airport code
ANIMAL	Animal
AT_VAT_NUMBER	Austrian VAT number
BG_VAT_NUMBER	Bulgarian VAT number
FR_VAT_NUMBER	French VAT number
BANK_ROUTING_TRANSIT_NUMBER	Bank routing transit number
BE_POSTAL_CODE	Belgian postal code
DE_POSTAL_CODE	German postal code
FR_POSTAL_CODE	French postal code

Semantic type	Description
UK_POSTAL_CODE	UK postal code
US_POSTAL_CODE	US postal code
BEVERAGE	Type of beverage
BOOLEAN	Answers with the value True or False
CA_PROVINCE_TERRITORY	Canadian province
CA_PROVINCE_TERRITORY_CODE	Canadian province code
CITY	City name
CIVILITY	Civility
COLOR_HEX_CODE	Color hexadecimal code
COMPANY	Company name
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
DECIMAL	Decimal numeric value
DE_PHONE	German phone number
FR_PHONE	French phone number
UK_PHONE	UK phone number
US_PHONE	US phone number
PHONE	Phone number. Includes german, french, british and american phone numbers.
DATE	Date including day, month and year
EMAIL	Email address
EN_MONEY_AMOUNT	Amount of money in English format

Semantic type	Description
FR_MONEY_AMOUNT	Amount of money in French format
EN_MONTH	Month in English
EN_MONTH_ABBREV	English month abbreviation
EN_WEEK_DAY	Week day or their abbreviation
FIRST_NAME	First name
LAST_NAME	Last name
FULL_NAME	Full name
FR_CODE_COMMUNE_INSEE	French Insee code of cities with Corsica and colonies
FR_COMMUNE	French municipality
FR_DEPARTEMENT	French department
FR_REGION	French region
FR_REGION_LEGACY	Former French regions
FR_SSN	French social security number
SE_SSN	Swedish person number
UK_SSN	National identification number, national identity number, or national insurance number generally called NI number
US_SSN	US social security number
GENDER	Gender
GPS_COORDINATE	Google Maps style GPS decimal format
HR_DEPARTMENT	HR department
INDUSTRY	Industry name
INDUSTRY_GROUP	Industry group
INTEGER	Numeric value
IBAN	International Bank Account Number
IPv4_ADDRESS	IPv4 address
IPv6_ADDRESS	IPv6 address

Semantic type	Description
ISBN_10	International standard book number 10 digits
ISBN_13	International standard book number 13 digits
JOB_TITLE	Job title
LANGUAGE	Language
LANGUAGE_CODE_ISO2	2-letter language code
LANGUAGE_CODE_ISO3	3-letter language code
LOCALIZATION	Localization
LOCATION_COORDINATE	Latitude and longitude coordinates separated by a comma in the form: N 0:59:59.99,E 0:59:59.99
MAC_ADDRESS	MAC address
MASTERCARD	Mastercard credit card
MEASURE_UNIT	Measure unit
MONTH	Month
MUSEUM	Museum name
MX_ESTADO	Mexican state
MX_ESTADO_CODE	Mexican state code
NA_STATE	North American states. Includes Mexican, Canadian and American states.
NA_STATE_CODE	North American state codes. Includes Mexican, Canadian and American state codes.
ORGANIZATION	Organization
PASSPORT	Passport number
SECTOR	Sector
SEDOL	Stock exchange daily official list
STREET_TYPE	Street type
TEXT	String text
URL	Web site URL
US_COUNTY	US county name

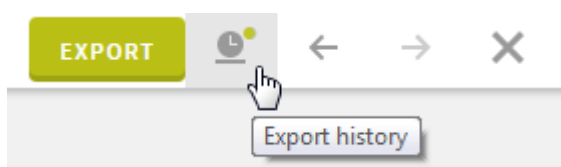
Semantic type	Description
US_CREDIT_CARD	US American Express credit card
US_STATE	US states
US_STATE_CODE	US state code
VISACARD	Visa credit card
WEB_DOMAIN	Web site domain
WEEKDAY	Day of the week

The export history page

Exports made on datasets larger than 10,000 rows by default are kept in memory. This makes it easier to retrieve the results of a given preparation with its specific export settings.

For a given preparation, a full history of the exports that have been performed is available.

To access the history of your preparation, click the **Export History** button in the application header bar.



The **Export History** page opens, where various information on the export process is displayed.

At first glance, you can get basic information on an export.

✓ SUCCESSFUL	Nicolas Talend	2017-04-11 16:40:54	2017-04-11 16:41:15	Talend Data Preparation	↓	CSV	▼
--------------	----------------	---------------------	---------------------	-------------------------	---	-----	---

Field	Description
Status	Describes the status of the export: in progress , successful , failed or canceled .
User	Shows which user made the export.
Start and end date	Describes the date and time at which the export process was launched, and then finished.
Execution process	Describes on which execution server the export was launched: the Talend Data Preparation server, or the Hadoop cluster.

Field	Description
Download	Use this button to retrieve the result of the preparation.
Output format	Describes the output format selected for the export.

When you click on a specific export, additional information is displayed.

✓ SUCCESSFUL	Nicolas Talend	2017-04-11 16:40:54	2017-04-11 16:41:15	Talend Data Preparation	↓	CSV	^
Start time	2017-04-11 16:40:54						
End time	2017-04-11 16:41:15						
Duration	20 seconds						
Runtime	Talend Data Preparation						

Field	Description
Duration	Describes the total time needed to complete the export process.
Runtime	Describes on which execution server the export was launched: the Talend Data Preparation server, or the Hadoop cluster.

User Roles

When managing authorized Talend Data Preparation users in Talend Administration Center, you have the possibility to apply predefined roles to them. Each role gives the rights to different actions.

The table below describes the specific roles/rights for the different types of Talend Data Preparation users. All the other actions are available for every user type.

	Administrator	Dataset Manager	Data Preparator
Add, edit or remove live dataset	✓	✗	✗
Certify dataset	✓	✗	✗
Export complete dataset	✓	✓	✗
Add,edit or remove a "database" dataset	✓	✓	✗

	Administrator	Dataset Manager	Data Preparator
Add, edit or remove a CSV, Parquet, Avro or JSON dataset on HDFS	✓	✓	✗
Export a preparations as CSV, Parquet, Avro or JSON dataset on HDFS	✓	✓	✗
Add,edit or remove a Salesforce dataset	✓	✓	✗
Add,edit or remove an Amazon S3 dataset	✓	✓	✗

Using metric prefixes

In order to write very large or very small numbers, you can use symbols from the metric system, such as 1k for 1000.

Symbols	Name	Description
T	Tera	1T = 1000000000000
G	Giga	1G = 1000000000
M	Mega	1M = 1000000
k	Kilo	1k = 1000
h	Hecto	1h = 100
da	Deca	1da = 10
d	Deci	1d = 0.1
c	Centi	1c = 0.01
m	Milli	1m = 0.001
μ	Micro	1μ = 0.000001
n	Nano	1n = 0.000000001
p	Pico	1p = 0.000000000001