Connecting the
Data-Driven Enterprise

Talend Data Preparation Quick Examples

2.1

Contents

Copyright....................3

Consolidating a list of phone numbers coming from a CRM solution....................4

Recreating email addresses before uploading them to a marketing solution....................10

Cleansing data coming from a human resource management system....................18

Preparing client data to upload it to a marketing solution....................21

Preparing an HDFS-based dataset....................25

Copyright

Adapted for 2.1. Supersedes previous releases.

Publication date: June 29th, 2017

Notices

End User License Agreement

Consolidating a list of phone numbers coming from a CRM solution

The **CRM Export** dataset corresponds to an Excel file that has been exported from a CRM solution.

It contains a list of people with their phone numbers for both regular and mobile phones. As these phone numbers are French, they are 10 digits long; numbers starting with 01, 02, 03, 04, or 05 correspond to landline numbers and numbers starting with 06 correspond to mobile phone numbers.

In order to have a consolidated list of phone numbers to call, you will create a new column listing all the mobile phone numbers: this column will also be filled with the landline number when there is no mobile phone. Also, if the landline and mobile phone numbers have been mixed up, you will correct them.

Whereas in other tools you would use conditions like "if" to perform these actions, with Talend Data Preparation, you will create filters.

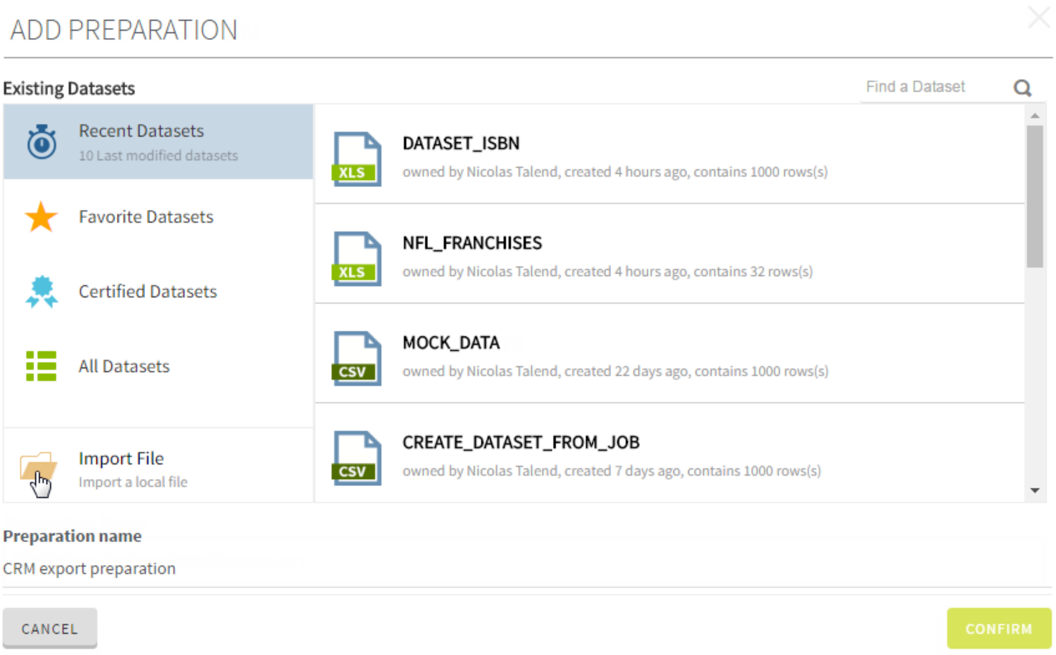
Retrieve the `CRM_export.xlsx` file from the **Downloads** tab of the online version of this page, at https://help.talend.com.

Adding a preparation for the **CRM Export** dataset

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. From the homepage, 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, `CRM export preparation` in this example.
4. Click **Import File** and select `CRM_export.csv` to use this file as dataset.



DATASET_ISBN
owned by Nicolas Talend, created 4 hours ago, contains 1000 rows(s)

NFL_FRANCHISES
owned by Nicolas Talend, created 4 hours ago, contains 32 rows(s)

MOCK_DATA
owned by Nicolas Talend, created 22 days ago, contains 1000 rows(s)

CREATE_DATASET_FROM_JOB
owned by Nicolas Talend, created 7 days ago, contains 1000 rows(s)

Preparation name
CRM export preparation

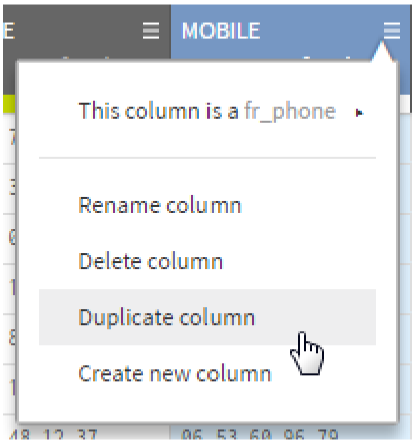
Your dataset opens with an empty recipe, and you can start adding preparation steps. All your changes are automatically saved.

Duplicating a column

In order to have a copy of the original data without having to create a new column and manually copy the data, you can simply duplicate a column.

Before working on the data, you will create a new column to receive the consolidated numbers.

1. Click the column you want to duplicate, **Mobile** in this example.
2. Click the white arrow or right-click the column to open the contextual menu.
3. Click **Duplicate Column**.



   A copy of the column, **Mobile_Copy**, is created with the same data as the original.

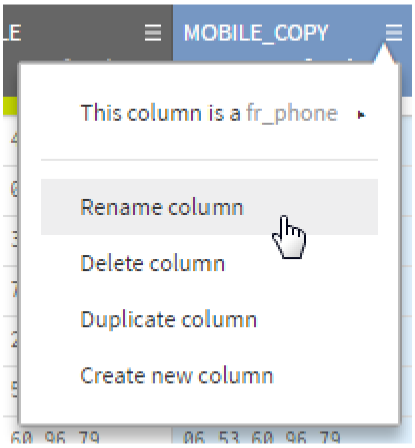
The **Mobile** column is duplicated and can now receive the consolidated data.

Renaming a column

In order to better identify a column, you can rename it.

You will rename the column previously created and give it a meaningful name.

1. Click the column you want to rename, **Mobile_Copy** in this example.
2. Perform one of the following actions:
   - Double-click the name of the column.
   - Use the **Rename Column** option in the in the contextual menu for the column.



3. Enter `Phone number to use` as new name for the selected column and press **Enter** to apply it.

The **Mobile_Copy** column is renamed.

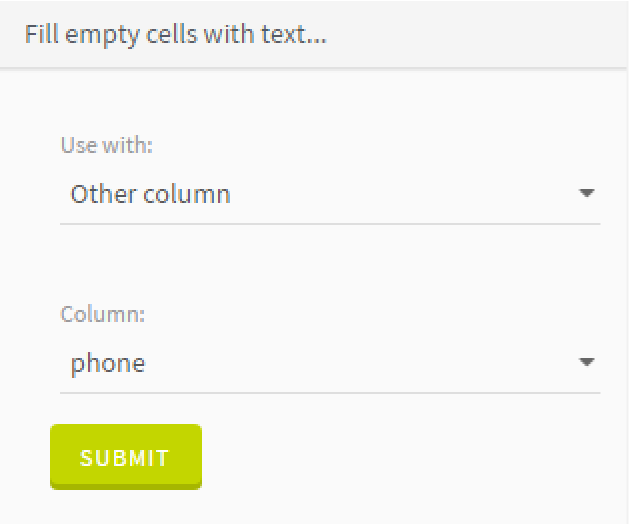
Filling empty cells with data from another column

The white part of the quality bar indicates that the column contains empty records.

You will fill the empty cells from the **Phone number to use**, with landline numbers extracted from the **Phone** column.

1. Select the **Phone number to use** column.

   You can identify columns containing empty records using the quality bar. The white part of it indicates that the column contains empty records. Data that matches the column type is shown in green, while orange shows 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. Configure the function as follows.



Fill empty cells with text...

Use with: Other column

Column: phone

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

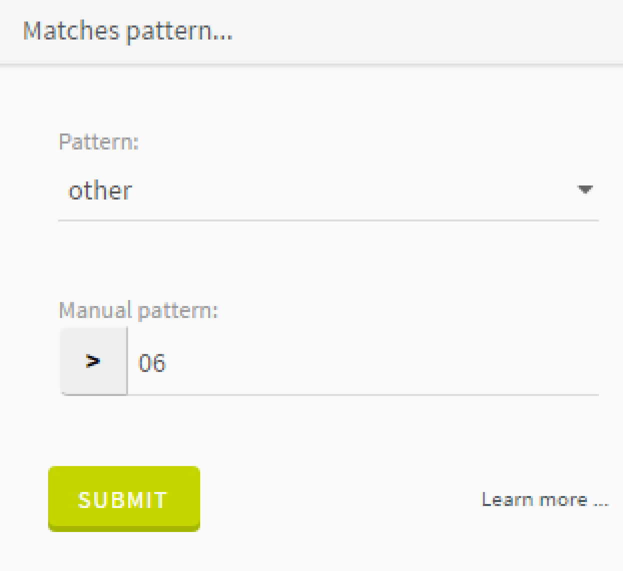
The empty cells in the **Phone number to use** column are filled with data from the **Phone** column.

Identifying if your data matches a pattern

If you want to identify whether there are mobile phone numbers among the landline numbers, you can use the **Matches Pattern** function.

Because the French mobile phone numbers start with 06, you are going to search for this pattern in the **Phone** column.

1. Select the column on which you want to apply the pattern, **Phone** in this example.
2. In the functions panel, type `Matches Pattern` and click the result to open the options for the associated function.
3. Configure the function as follows.



Matches pattern...

Pattern: other

Manual pattern: 06

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

   A new column is created, with the value **true** if the pattern matches and **false** if it does not

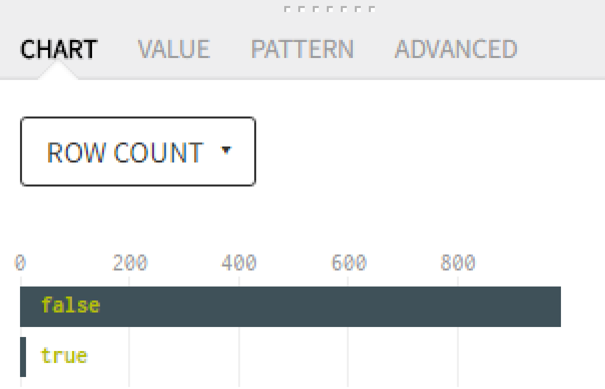
The information is stored in a new column and you can now apply a filter to isolate the mobile phone numbers present in the **Phone** column.

Creating a filter

Creating a filter is a quick way to identify or isolate data.

You will isolate the values in the **Phone** column that actually match the pattern defined in the previous step. This will be possible through the use of a filter on the values that were found as **true** in the **Phone_Matching** column.

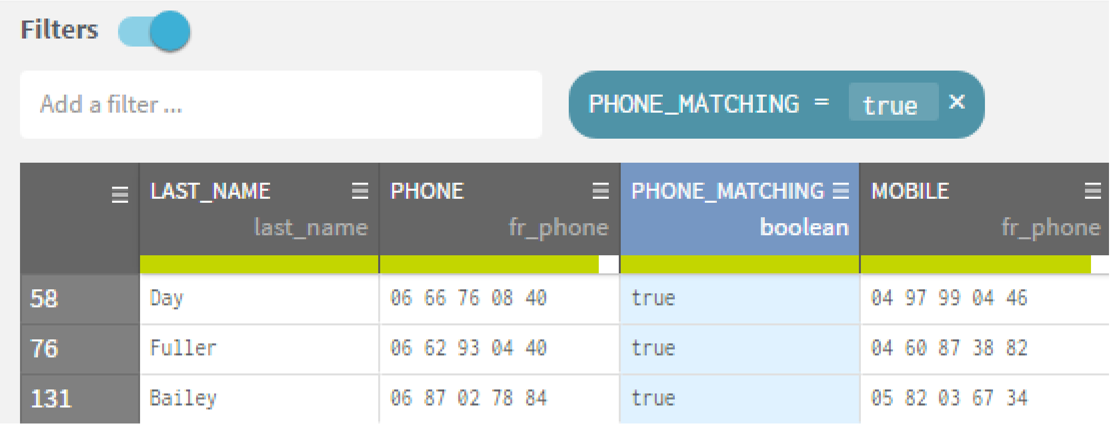
1. Select the **Phone_Matching** column.
2. In the data profiling area, on the bottom right of the screen, you can see an horizontal bar chart, displaying the number of occurences of the **true** and **false** values.



   Diagrams are a convenient and quick way to apply filters on your data, for one or several columns at a time.

3. Click the bar displaying the **true** values.

   A filter is applied on your data, and only the entries that match the value **true** are displayed in the grid, as you can see in the filter bar.



Filters

Add a filter ...

PHONE_MATCHING = true ×

	LAST_NAME last_name	PHONE fr_phone	PHONE_MATCHING boolean	MOBILE fr_phone
58	Day	06 66 76 08 40	true	04 97 99 04 46
76	Fuller	06 62 93 04 40	true	04 60 87 38 82
131	Bailey	06 87 02 78 84	true	05 82 03 67 34

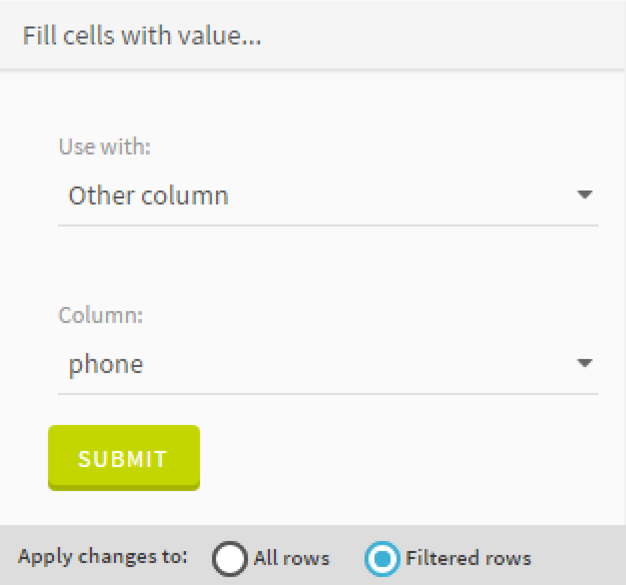
The data is filtered on the value **true**, and you will now work on this small sample of data.

Replacing values with data from another column

You can take the data from any column, and use it to consolidate another.

Now that only the mobile phone numbers are selected, you will use these values to replace the wrong numbers in the **Phone number to use** column.

1. Select a column in which you want to replace values, **Phone number to use** in this example.
2. In the **Functions panel**, type `Fill Cells with Value` and click the result to open the options for the associated function.
3. Configure the function as follows.



Fill cells with value...

Use with: Other column

Column: phone

Apply changes to: ○ All rows ● Filtered rows

Make sure the **Filtered Rows** option is selected in order to only apply the function on the filtered lines.

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

   The filtered content of the selected column, **Phone number to use** in this example, is replaced with the data from the **Phone** column.

5. In the filter bar, click the garbage bin icon to clear the filter and display the whole dataset again.

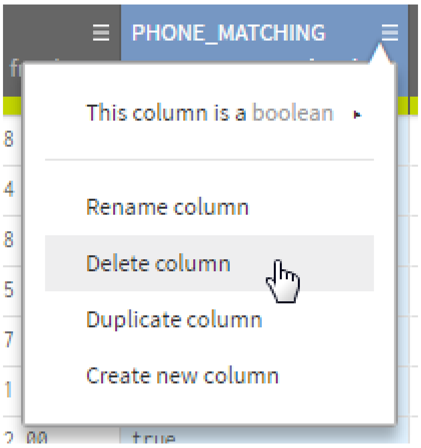
You have used the filtered data to update the information of the consolidated column.

Deleting a column

If you want to remove a column you do not need, you can delete a column.

Now that you have used the **Phone_Matching** for filtering purposes, you can delete it.

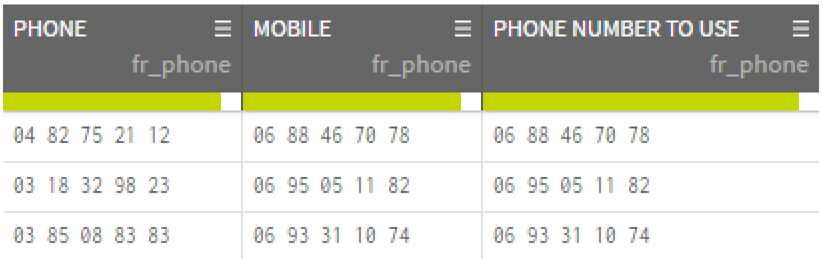
1. Click the column you want to delete, **Phone_Matching** in this example.
2. Click the white arrow or right-click the column to open the contextual menu.
3. Click **Delete Column**.



The **Phone_Matching** column is deleted.

Removing empty records from a column

The white part of the quality bar indicates that a column contains empty records. You may want to remove the rows containing these empty records.

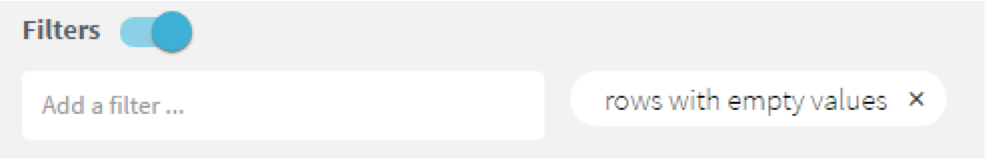


PHONE	MOBILE	PHONE NUMBER TO USE
fr_phone	fr_phone	fr_phone
04 82 75 21 12	06 88 46 70 78	06 88 46 70 78
03 18 32 98 23	06 95 05 11 82	06 95 05 11 82
03 85 08 83 83	06 93 31 10 74	06 93 31 10 74

In the quality bar, data that matches the column type is shown in green, while orange shows invalid data that does not match the column type.

1. In the top left of the grid, click the white arrow and select **Display rows with empty values**.

   This action applies a filter on all the empty entries from the dataset.



2. In the functions panel, on the top right of your screen, type `Delete these filtered rows` and click the result to execute the associated function.

   All the rows containing empty cells are removed from the dataset.
3. Click the bin icon in the filter bar to clear the filter and display the whole dataset again.

All the rows containing empty records are removed from the dataset and the quality bar under each column is now fully green.

Exporting the results of your preparation

Once your preparation is complete, you may want to export the data you have cleansed.

You have managed to filter and consolidate a list of phone numbers, and you can now export the result of this preparation.

1. Click the **Export** button.
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.

The data you cleansed using your preparation is exported to a local file.

Recreating email addresses before uploading them to a marketing solution

The **Marketing leads** dataset represents a file you received from a marketing campaign but the email addresses are missing.

In order for this file to be uploaded into your marketing solution, you have to create those email addresses. You will guess them from the name and the company of the prospects, and from the email format usually used by those companies.

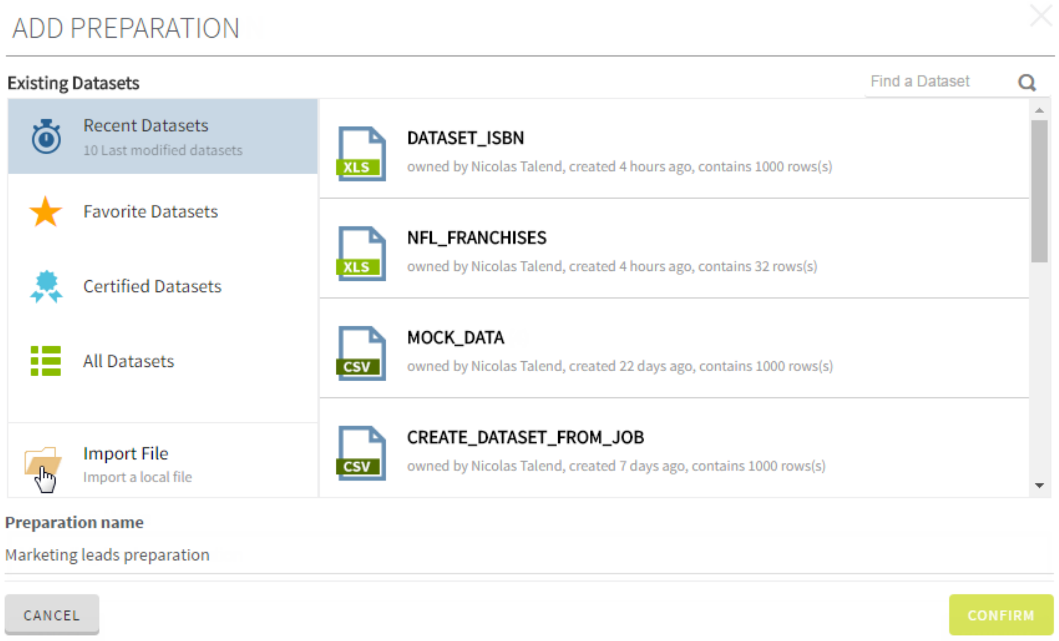
Retrieve the `marketing_leads.csv` and `emails_reference.csv` files from the **Downloads** tab of the online version of this page, at https://help.talend.com.

Adding a preparation for the **Marketing leads dataset**

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 be automatically saved in the preparations list, and all the changes you make are also saved automatically.

1. From the homepage, 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, `Marketing leads preparation` in this example.
4. Click **Import File** and select `marketing_leads.csv` to use this file as dataset.



DATASET_ISBN
owned by Nicolas Talend, created 4 hours ago, contains 1000 rows(s)

NFL_FRANCHISES
owned by Nicolas Talend, created 4 hours ago, contains 32 rows(s)

MOCK_DATA
owned by Nicolas Talend, created 22 days ago, contains 1000 rows(s)

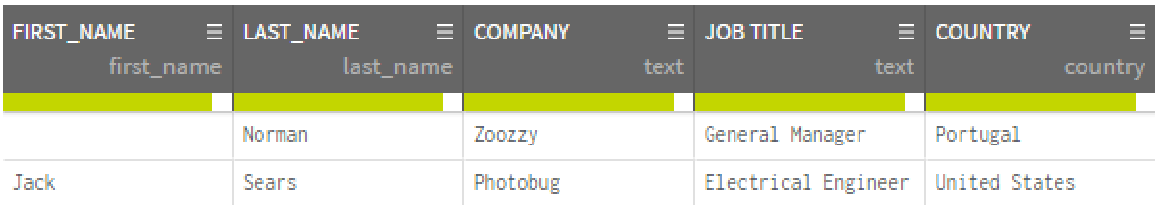
CREATE_DATASET_FROM_JOB
owned by Nicolas Talend, created 7 days ago, contains 1000 rows(s)

Preparation name: Marketing leads preparation

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

Removing empty records from the dataset

The white part of the quality bar indicates that a column contains empty records. You may want to remove the rows containing these empty records.



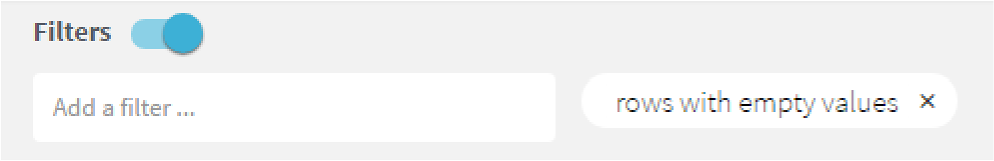
FIRST_NAME first_name	LAST_NAME last_name	COMPANY text	JOB TITLE text	COUNTRY country
	Norman	Zoozzy	General Manager	Portugal
Jack	Sears	Photobug	Electrical Engineer	United States

In the quality bar, data that matches the column type is shown in green, while orange shows invalid data that does not match the column type.

1. In the **First_Name** column, click the white part of the quality bar.
2. Click **Delete the Rows with Empty Cell** to remove the rows with missing data.

   You can perform this action for any given column, but there is a simpler way to remove all the empty rows from your dataset.
3. In the top left of the grid, click the white arrow and select **Display rows with empty values**.

   This action applies a filter on all the empty entries from the dataset.



4. In the functions panel, on the top right of your screen, type `Delete these filtered rows` and click the result to execute the associated function.

   The remaining rows containing empty cells are removed from the dataset.
5. Click the bin icon in the filter bar to clear the filter and display the whole dataset again.

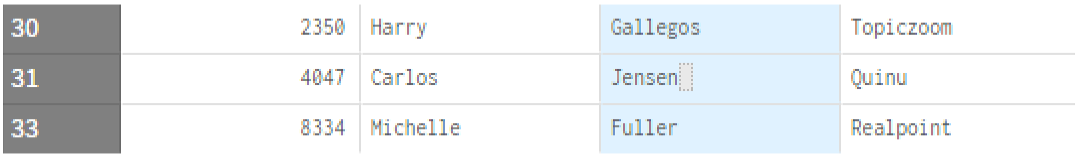
All the rows containing empty records are removed from the dataset and the quality bar under each column is now fully green.

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.

You can see that the **First_Name** and **Last_Name**columns contain some entries with blank spaces.



30	2350	Harry	Gallegos	Topiczoom
31	4047	Carlos	Jensen	Quinu
33	8334	Michelle	Fuller	Realpoint

1. Select the **First_Name** column.
2. While keeping the **Ctrl** button pressed, click the header of the **Last_Name** column.

   The two columns are now selected, and you can apply a function to both columns in one action.
3. In the functions panel, type `Remove trailing and leading characters` and click the result to open the options for the associated function.
4. In the **Padding character** drop-down list, select **whitespace** and click **Submit**.

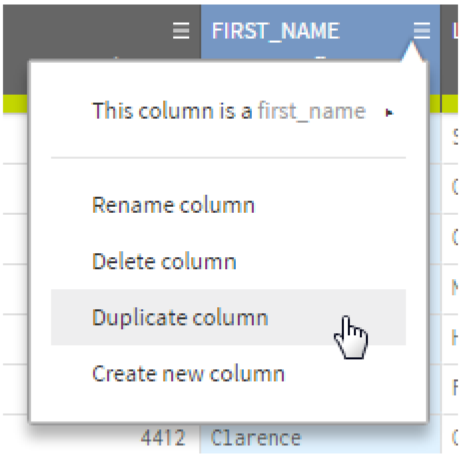
Blank spaces are removed from the selected columns.

Duplicating columns

If you want to have a copy of a given column, you may want to duplicate that column.

You will duplicate two columns and use them as a base for recreating the email addresses.

1. Click the column you want to duplicate, **First_Name** in this example.
2. Click the white arrow or right-click the column to open the contextual menu.
3. Click **Duplicate Column**.



A copy of the column, **First_Name_Copy**, is created with the same data as the original.

4. Repeat these actions to duplicate the **Last_Name** column.

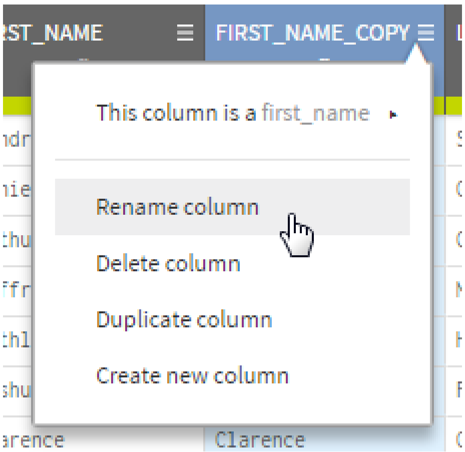
The two columns are duplicated. The information contained in those two column will be used to create the local part of the email addresses you want to create.

Renaming columns

In order to better identify a column, you can rename it.

You will rename the two columns previously created and give them a meaningful name. Those new colums will be used to create the local part of the email addresses.

1. Click the column you want to rename, **First_Name_Copy** in this example.
2. Perform one of the following actions:
   - Double-click the name of the column.
   - Use the **Rename Column** option in the column contextual menu.



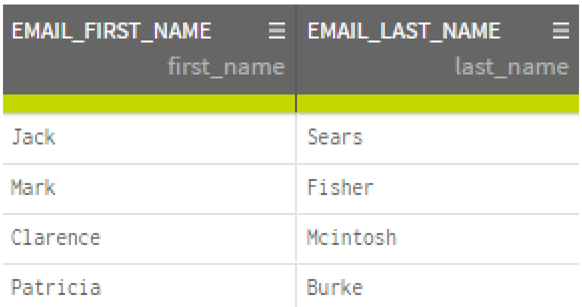
3. Enter `Email_First_Name` as new column name and press **Enter** to apply it.
4. Repeat these steps to rename the **Last_Name_Copy** column as **Email_Last_Name**.

The selected columns are renamed.

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.

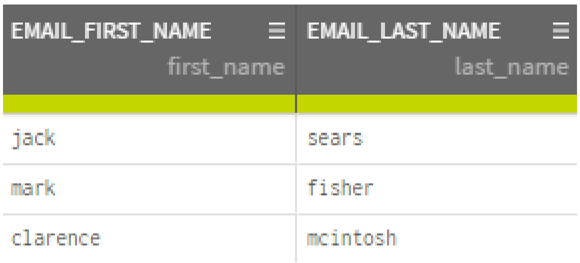
The data contained in the two renamed columns starts with an upper case. You will put the all the text in lower case, in order to merge the two columns.



EMAIL_FIRST_NAME	EMAIL_LAST_NAME
first_name	last_name
Jack	Sears
Mark	Fisher
Clarence	Mcintosh
Patricia	Burke

1. Select the column containing text you want to change to lower case, **Email_First_Name** in this example.
2. In the functions panel, type `Change Style to lower Case` and click the result to execute the associated function.
3. Repeat these steps to put the **Email_Last_Name** column in lower case.

The text contained in the two colums is now in lower case.



EMAIL_FIRST_NAME first_name	EMAIL_LAST_NAME last_name
jack	sears
mark	fisher
clarence	mcintosh

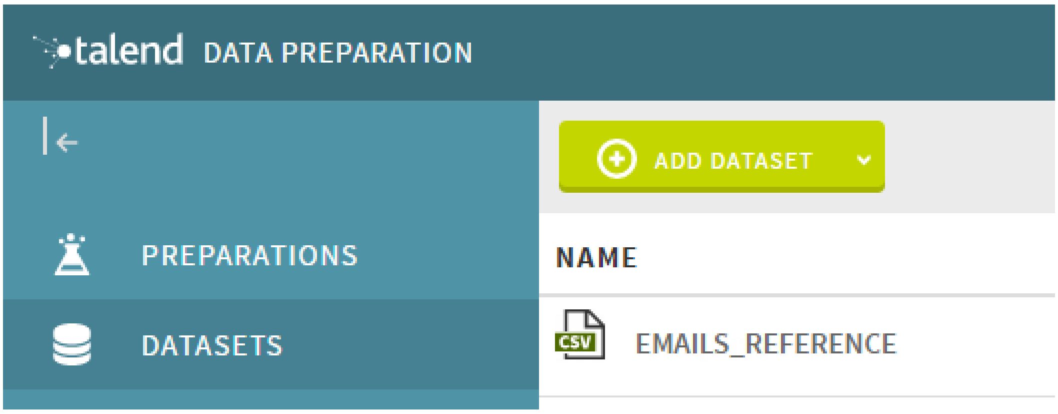
Dynamically using the data from another dataset

The lookup feature matches data from the current dataset with its counterpart in a reference dataset.

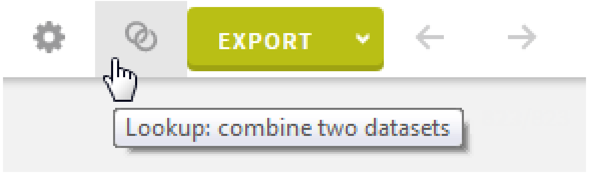
On the one hand, you have the `marketing_leads` dataset, that you are currently working on, that contains information about the company where the listed customers are working. On the other hand, the `emails_reference` contains a list of companies, and the email domain that they are using.

You are going to do a lookup on the `emails_reference` dataset, and extract the information about email domains to match them with the companies from the `marketing_leads` dataset.

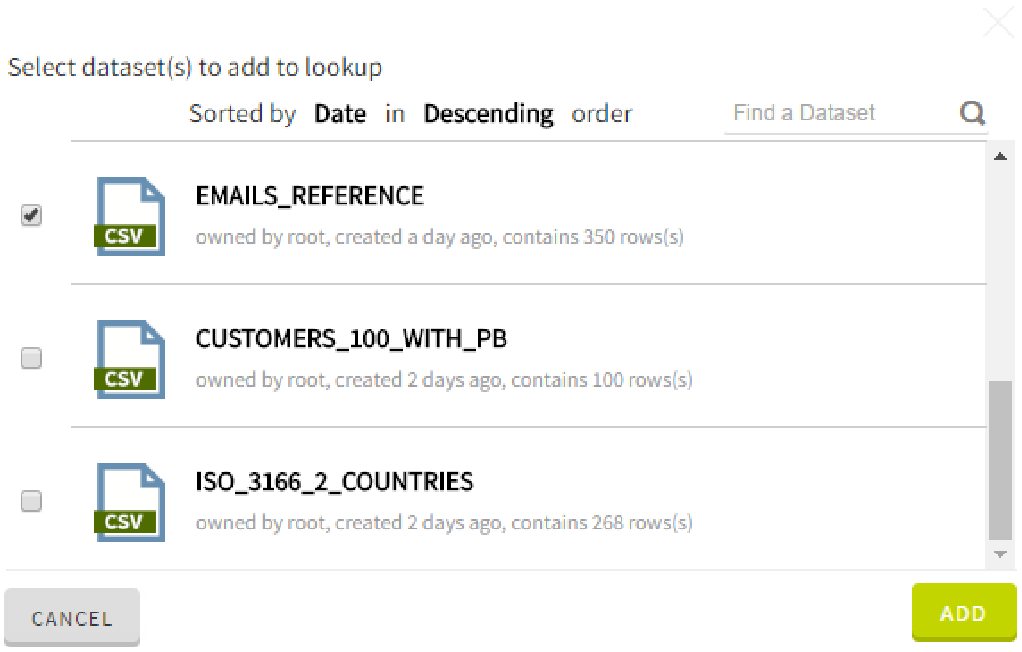
To perform the lookup on the `emails_reference`, you have to import it by using the **Add dataset** button in the **Datasets** view of the homepage.



1. Select the column on which you want to perform the lookup, the **Company** column in this example. This is the column that can be found in the source dataset, as well as the reference dataset. There must always be a common column between two datasets to perform a lookup.
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, the **Emails_Reference** dataset in this example.



Select dataset(s) to add to lookup

Sorted by **Date** in **Descending** order

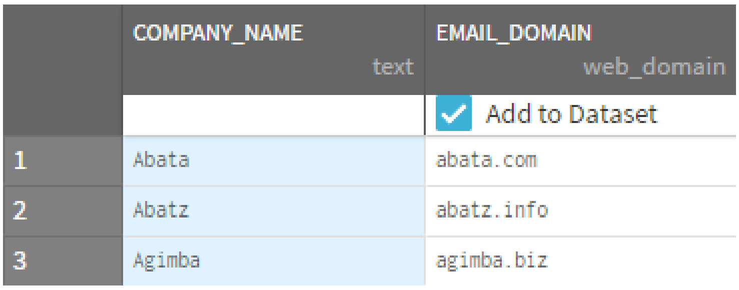
Find a Dataset

☑ EMAILS_REFERENCE
owned by root, created a day ago, contains 350 rows(s)

☐ CUSTOMERS_100_WITH_PB
owned by root, created 2 days ago, contains 100 rows(s)

☐ ISO_3166_2_COUNTRIES
owned by root, created 2 days ago, contains 268 rows(s)

4. Click **Add**.
5. In the lookup window that opens in the bottom half of your screen, click the **Company_Name** column.
6. Select the **Add to Dataset** check box.



	COMPANY_NAME text	EMAIL_DOMAIN web_domain
		☑ Add to Dataset
1	Abata	abata.com
2	Abatz	abatz.info
3	Agimba	agimba.biz

7. Point your mouse over the **Confirm** button to preview the changes.
8. Click the **Confirm** button to apply those changes.

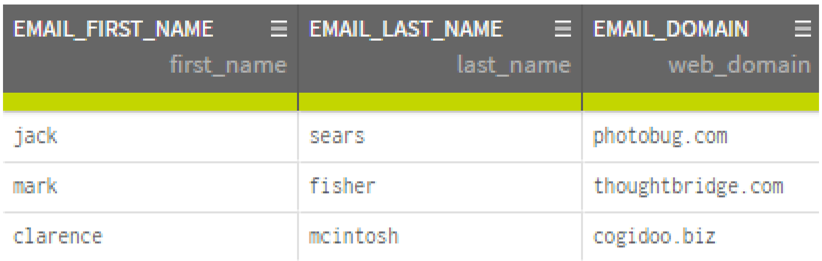
The **Email_Domain** column is added to the `marketing_leads` dataset, next to the **Company** column.

This information about email domains will be added to the first names and last names from the duplicated column to create the complete email addresses.

Merging the content of several columns

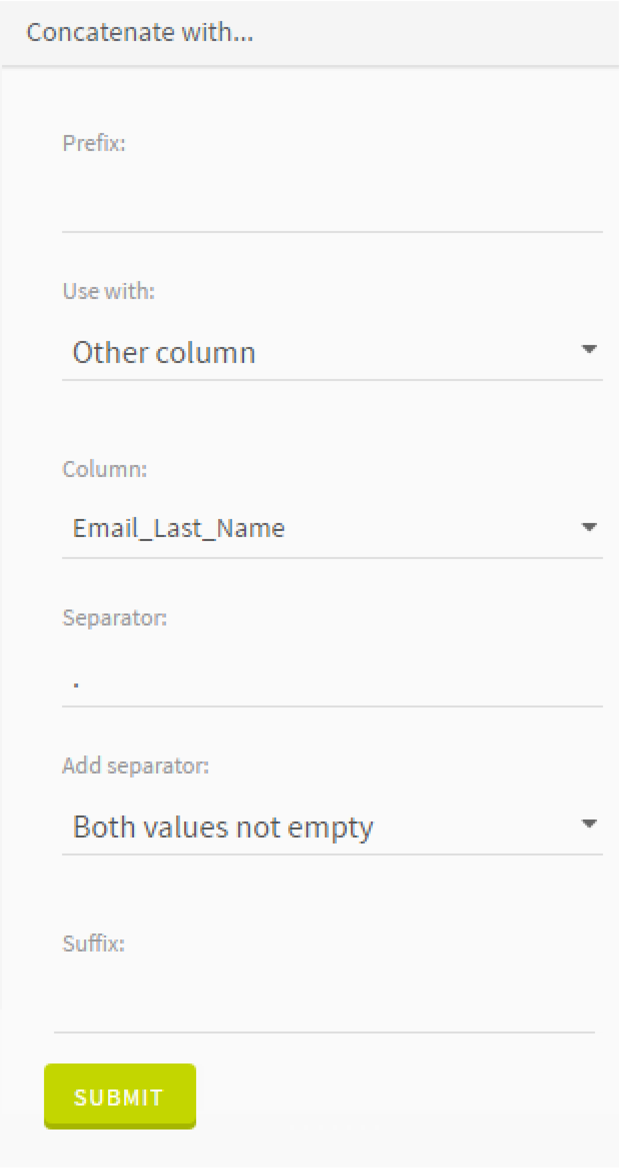
In some cases, the data you want to use is split in several columns. You can group these columns using a concatenation.

All the information you need to create the email addreses is now ready, and you only need to assemble it. You will merge the three columns that you have created since the beginning of this scenario.



EMAIL_FIRST_NAME first_name	EMAIL_LAST_NAME last_name	EMAIL_DOMAIN web_domain
jack	sears	photobug.com
mark	fisher	thoughtbridge.com
clarence	mcintosh	cogidoo.biz

1. Select the **Email_First_Name** column.
   When merging several columns together, the one that you select at the beginning, **Email_First_Name** in this case, 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 function as follows:



Concatenate with...

Prefix:

Use with: Other column

Column: Email_Last_Name

Separator: .

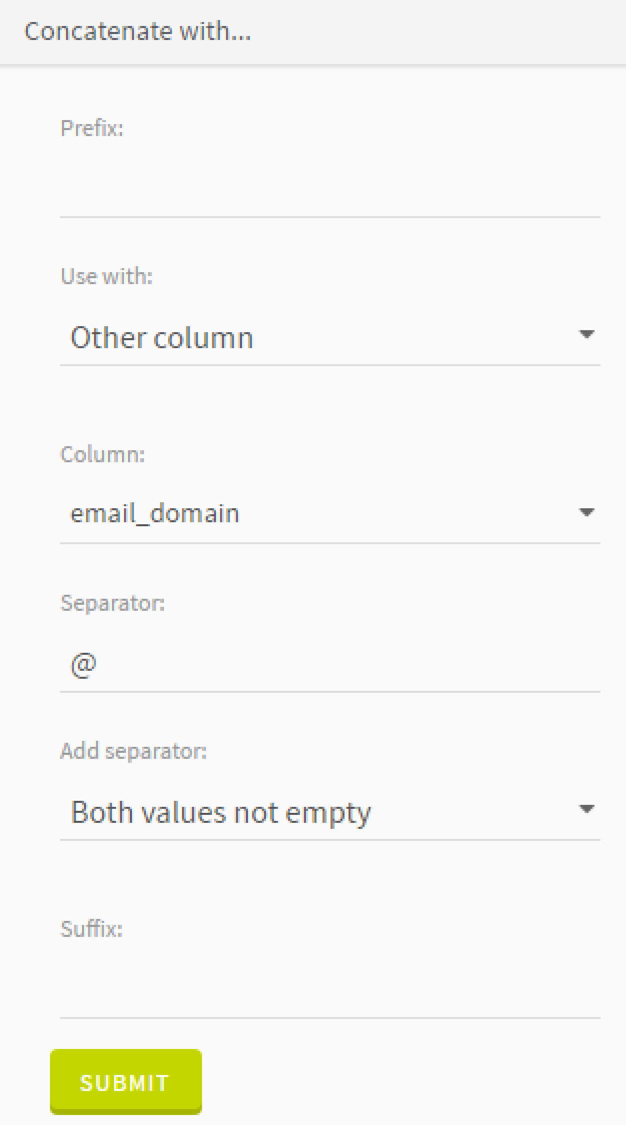
Add separator: Both values not empty

Suffix:

You can only concatenate two columns at a time, so you will begin by merging the **Email_First_Name** column with the **Email_Last_Name**, with **.** as separator.

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

   A new column with the merged content from the two columns is created.
5. Proceed the same way to merge the column you just created with the **Email_Domain** column, but using @ as separator this time.



Concatenate with...

Prefix:

Use with: Other column

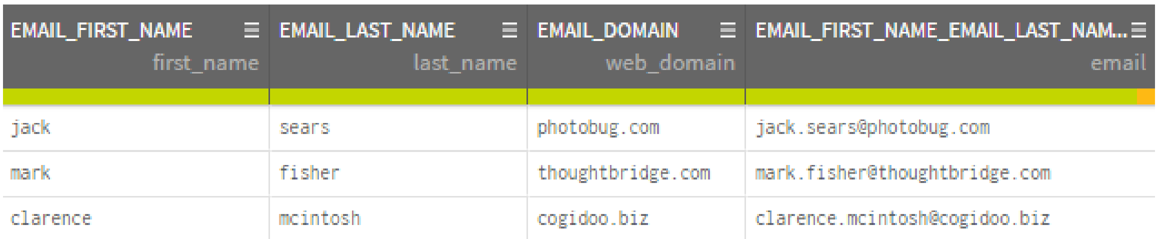
Column: email_domain

Separator: @

Add separator: Both values not empty

Suffix:

The content of the three columns has been merged. You have created a column containing valid email addresses, based on first names, last names and a web domain for each company.



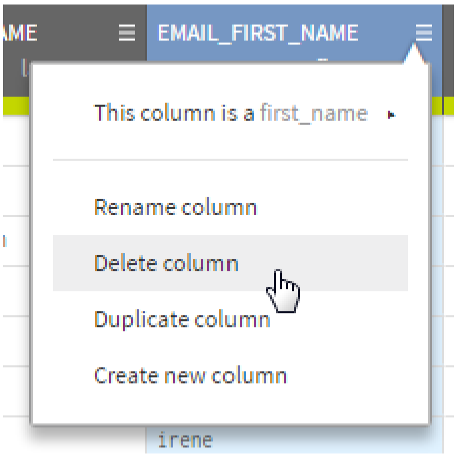
EMAIL_FIRST_NAME	EMAIL_LAST_NAME	EMAIL_DOMAIN	EMAIL_FIRST_NAME_EMAIL_LAST_NAM...
first_name	last_name	web_domain	email
jack	sears	photobug.com	jack.sears@photobug.com
mark	fisher	thoughtbridge.com	mark.fisher@thoughtbridge.com
clarence	mcintosh	cogidoo.biz	clarence.mcintosh@cogidoo.biz

Deleting a column

If you want to remove a column you do not need, you can delete a column.

Now that you have used the columns for the concatenation, you can delete them and only keep the result.

1. Click the column you want to delete, **Email_First_Name** for example.
2. Click the white arrow or right-click the column to open the contextual menu.
3. Click **Delete Column**.



4. Proceed the same way to delete the **Email_Last_Name** column and the **Email_Domain** column.

The selected columns are deleted.

Exporting the results of your preparation

Once your preparation is complete, you may want to export the data you have cleansed.

You have managed to prepare your dataset to recreate the information that was originally missing, the customers email addresses in this case.

1. Click the **Export** button.
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.

The data you cleansed using your preparation is exported to a local file.

Cleansing data coming from a human resource management system

The **HRMS Export** dataset corresponds to an Excel file that has been exported from an American human resource management system (or HRMS).

It contains the full list of employees since the creation of the company with their name, job title, hiring date, departure date if any and bank information for their salary. In this dataset, the dates are in the American date format and you want to transform them to the French date format so that they can be used with French software solutions. Also, you want to extract the account number from the IBAN number for the French accounts.

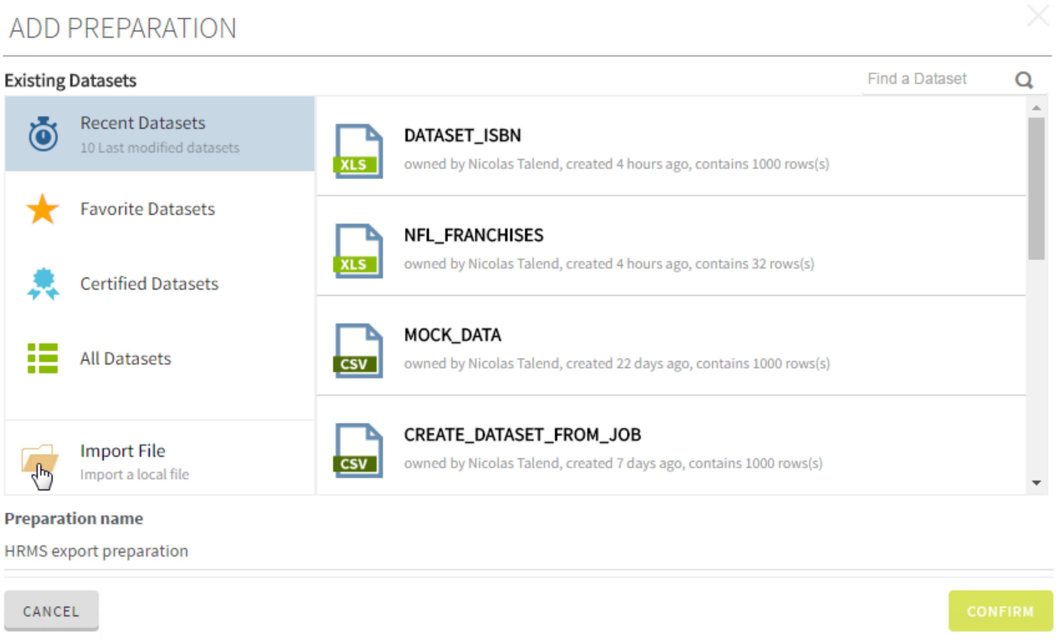
Retrieve the `HRMS_export.xlsx` file from the **Downloads** tab of the online version of this page, at https://help.talend.com.

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 be automatically saved in the preparations list, and all the changes you make are also saved automatically.

1. From the homepage, 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, `HRMS export preparation` for example.
4. Click **Import File** and select `HRMS_export.xlsx` to use this file as dataset.



DATASET_ISBN
owned by Nicolas Talend, created 4 hours ago, contains 1000 rows(s)

NFL_FRANCHISES
owned by Nicolas Talend, created 4 hours ago, contains 32 rows(s)

MOCK_DATA
owned by Nicolas Talend, created 22 days ago, contains 1000 rows(s)

CREATE_DATASET_FROM_JOB
owned by Nicolas Talend, created 7 days ago, contains 1000 rows(s)

Preparation name
HRMS export preparation

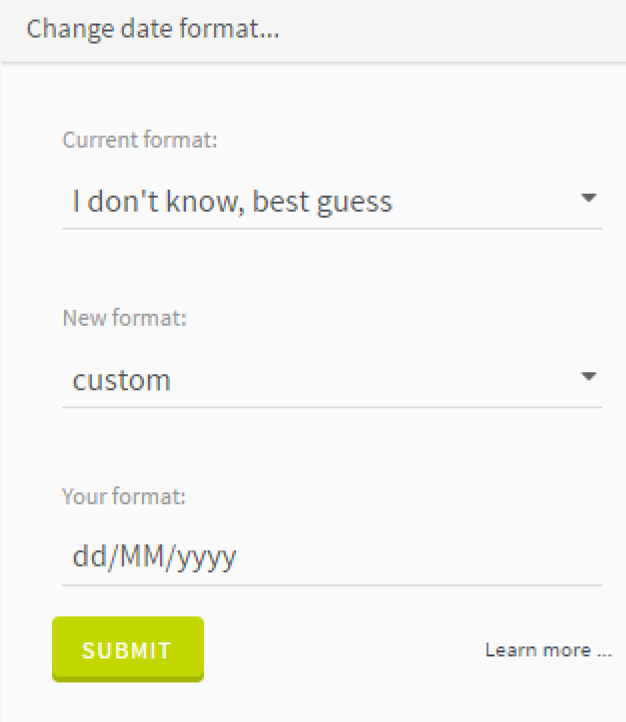
Your dataset opens with an empty recipe, and you can start adding preparation steps. All your changes are automatically saved.

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.

You will change the date format that is used in this dataset, from the American format, to the French format.

1. Select the **Entry_Date** column.
2. In the functions panel, on the right side of the screen, type `Change Date Format` and click the result to open the options for the associated function.
3. Configure the function as follows.



Change date format...

Current format: I don't know, best guess

New format: custom

Your format: dd/MM/yyyy

The French format that you want to use, `dd/MM/yyyy`, is not available by defaut so you have to enter it as a custom value in the **Your Format** field.

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

4. Point your mouse over the **Submit** button to preview the changes, and click it to apply the function.

The date format is changed in the selected column.

Extracting the bank account number

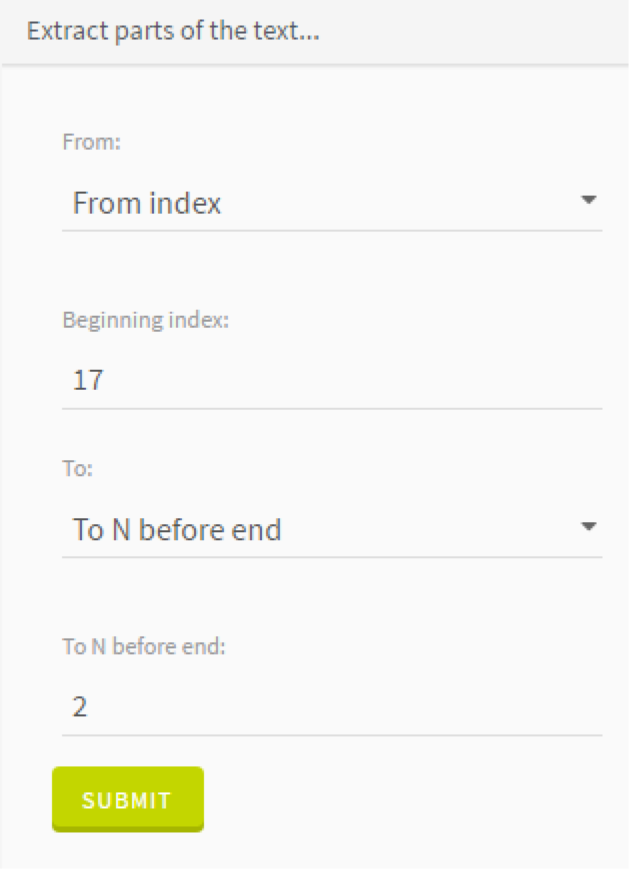
If you want to take part of the text contained in a cell and reuse it elsewhere, you can extract part of the text.

The **HRMS Export** preparation contains French International Bank Account Nnumbers (IBAN). An IBAN is a 33-characters code, including spaces. It is made of a Country code, two check digits, a five-digit bank identifier, a five-digit branch identifier, an eleven-digit account number, and two final check digits.

You will extract the account number part of those IBAN, to a new column.

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 the **IBAN** column.
2. In the functions panel, type `Extract Parts of the Text` and click the result to display the options of the associated function.
3. Configure the function as follows:



Extract parts of the text...

From:

From index

Beginning index:

17

To:

To N before end

To N before end:

2

The selection will start at the 17th character, spaces included, and end two characters before the end.

4. Click the **Submit** button to extract the selection you made to a new column, **IBAN_Substring** in this case.

The text corresponding to the selection you made is extracted to a new column, that you can rename if you want.

Exporting the results of your preparation

Once your preparation is complete, you may want to export the data you have cleansed.

The preparation on the `hrms_export.xlsx` aimed at changing the date format and extracting the account numbers from the IBAN, is now complete and you can export it.

1. Click the **Export** button.
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.

The data you cleansed using your preparation is exported to a local file.

Preparing client data to upload it to a marketing solution

The **Customer Contact Data** dataset represents a file containing a list of clients with different information such as their name, their company or their country.

You will prepare and clean this data in order to be able to upload it to a marketing solution.

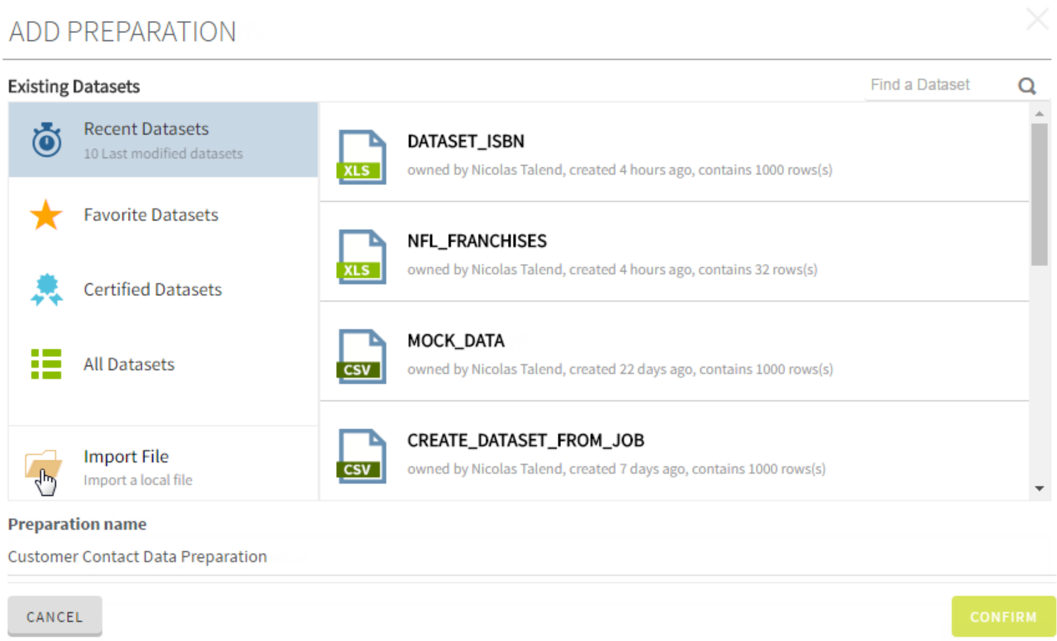
Retrieve the `customer_contact_data.csv` file from the **Downloads** tab of the online version of this page, at https://help.talend.com.

Adding a preparation for the **Customer Contact Data** dataset

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. From the homepage, 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, `Customer Contact Data Preparation` for example.
4. Click **Import File** and select `customer_contact_data.csv` to use this file as dataset.



DATASET_ISBN
owned by Nicolas Talend, created 4 hours ago, contains 1000 rows(s)

NFL_FRANCHISES
owned by Nicolas Talend, created 4 hours ago, contains 32 rows(s)

MOCK_DATA
owned by Nicolas Talend, created 22 days ago, contains 1000 rows(s)

CREATE_DATASET_FROM_JOB
owned by Nicolas Talend, created 7 days ago, contains 1000 rows(s)

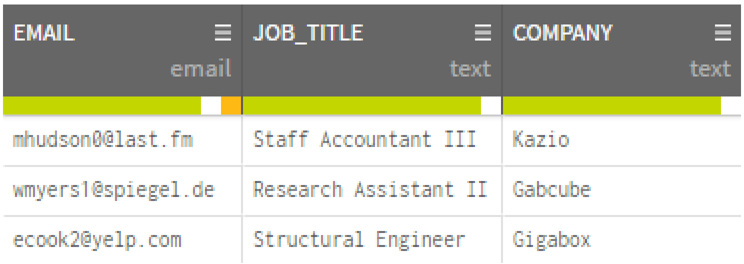
Preparation name
Customer Contact Data Preparation

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

Removing empty and invalid rows

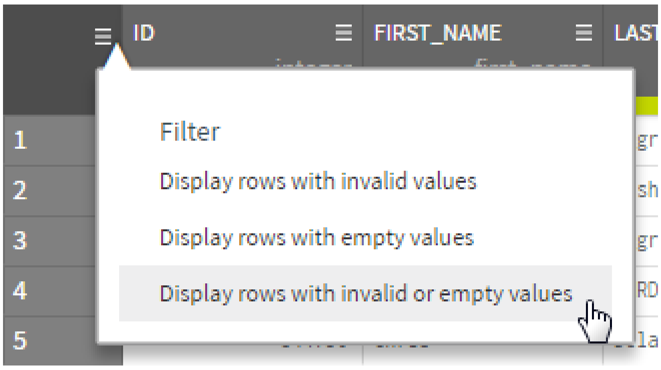
You can remove all the empty and invalid entries from a dataset in one go.

As you can see in the quality bar under each colum, the `customer_contact_data.csv` contains several rows with either empty or invalid cells. You are going to delete all these rows. Using the quality bar is a quick way of removing empty and invalid records for a given column, but you want to perform this on the whole dataset.



EMAIL email	JOB_TITLE text	COMPANY text
mhudson0@last.fm	Staff Accountant III	Kazio
wmyers1@spiegel.de	Research Assistant II	Gabcube
ecook2@yelp.com	Structural Engineer	Gigabox

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


ID
FIRST_NAME
LAST
1
2
3
4
5
Filter
Display rows with invalid values
Display rows with empty values
Display rows with invalid or empty values

You have actually applied a filter on your data, and only the empty and invalid values present in the dataset are displayed.

3. In the functions panel, type `Delete these filtered rows` and click the result to apply the associated function.

   Make sure that the **Filtered Rows** radio button is selected in front of the **Apply changes to** field.

   The rows containing empty or invalid entries are removed from the dataset.
4. Click the bin icon in the filter bar to clear the filter and display the whole dataset again.

All the rows containing empty records are removed from the dataset and the quality bar under each column is now fully green.

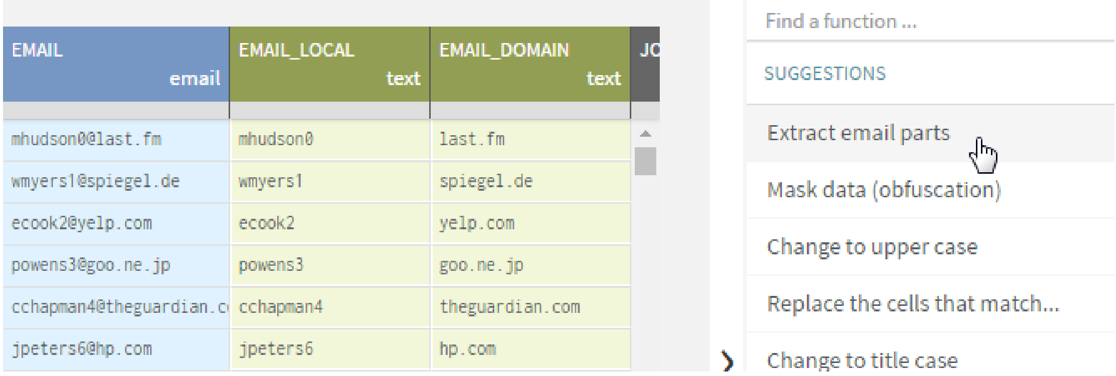
Extracting email address parts

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.

You will extract the two parts of the email addresses to make it easier to upload to the marketing solution.

1. Select the **Email** column.
2. In the functions panel, type `Extract Email Parts` and point your mouse over the function name to preview the changes.



EMAIL email	EMAIL_LOCAL text	EMAIL_DOMAIN text	JC
mhudson0@last.fm	mhudson0	last.fm	
wmyers1@spiegel.de	wmyers1	spiegel.de	
ecook2@yelp.com	ecook2	yelp.com	
powens3@goo.ne.jp	powens3	goo.ne.jp	
cchapman4@theguardian.c	cchapman4	theguardian.com	
jpeters6@hp.com	jpeters6	hp.com	

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

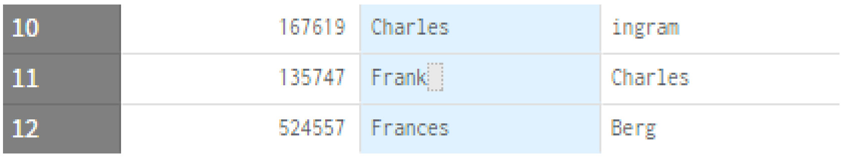
The local part and the domain part are extracted from the email addresses. The extracted data is put into two new columns.

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.

You can see that the **First_Name** column contains some entries with blank spaces.



10	167619	Charles	ingram
11	135747	Frank	Charles
12	524557	Frances	Berg

1. Select the **First_Name** column.
2. In the functions panel, type `Remove trailing and leading characters` and click the result to open the options for the associated function.
3. In the **Padding character** drop-down list, select **whitespace** and click **Submit**.
4. Repeat this action for every column containing blank spaces.

Blank spaces are removed from the selected column.

Exporting the results of your preparation

Once your preparation is complete, you may want to export the data you have cleansed.

You have cleaned the empty and invalid rows from your dataset, removed the unnecessary blank spaces, and extracted the information about the customers email adresses. The prepared dataset is now compatible with your marketing solution and you can export it.

1. Click the **Export** button.
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.

The data you cleansed using your preparation is exported to a local file.

Preparing an HDFS-based dataset

When using Talend Data Preparation in a big data context, you can access data stored on HDFS (Hadoop File System).

In this example, you work for a worldwide online video streaming company. You will retrieve some customer information stored on a cluster, create a dataset in Talend Data Preparation, apply various preparation steps to cleanse and enrich this data, and then export it back on the cluster with a new format.

Through the use of the Components Catalog service, the data is not physically stored on the Talend Data Preparation server, but rather fetched on-demand from the cluster. Only a sample is retrieved and display in the Talend Data Preparation interface for you to work on.

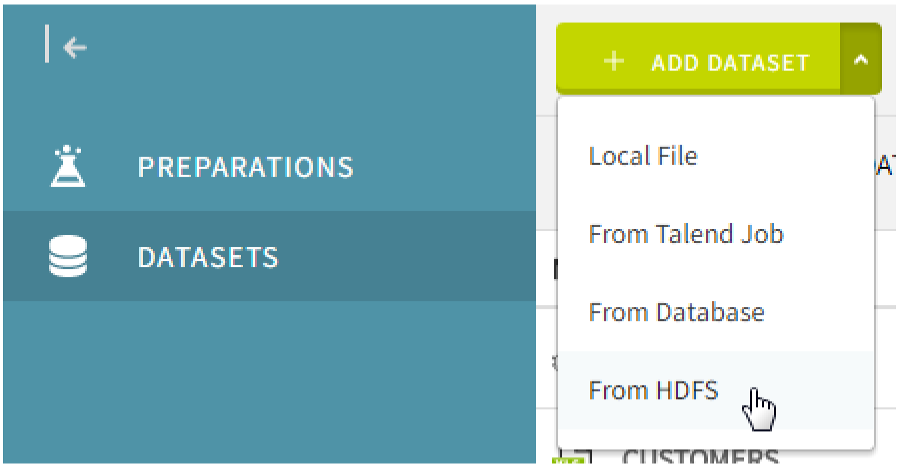
To use Talend Data Preparation in a Big Data context, you must fulfill the following prerequisites:

- The Components Catalog service is installed and running on a Windows or Linux machine.
- The Spark Job Server is installed and running on a Linux machine.
- The Streams Runner is installed and running on a Linux machine.

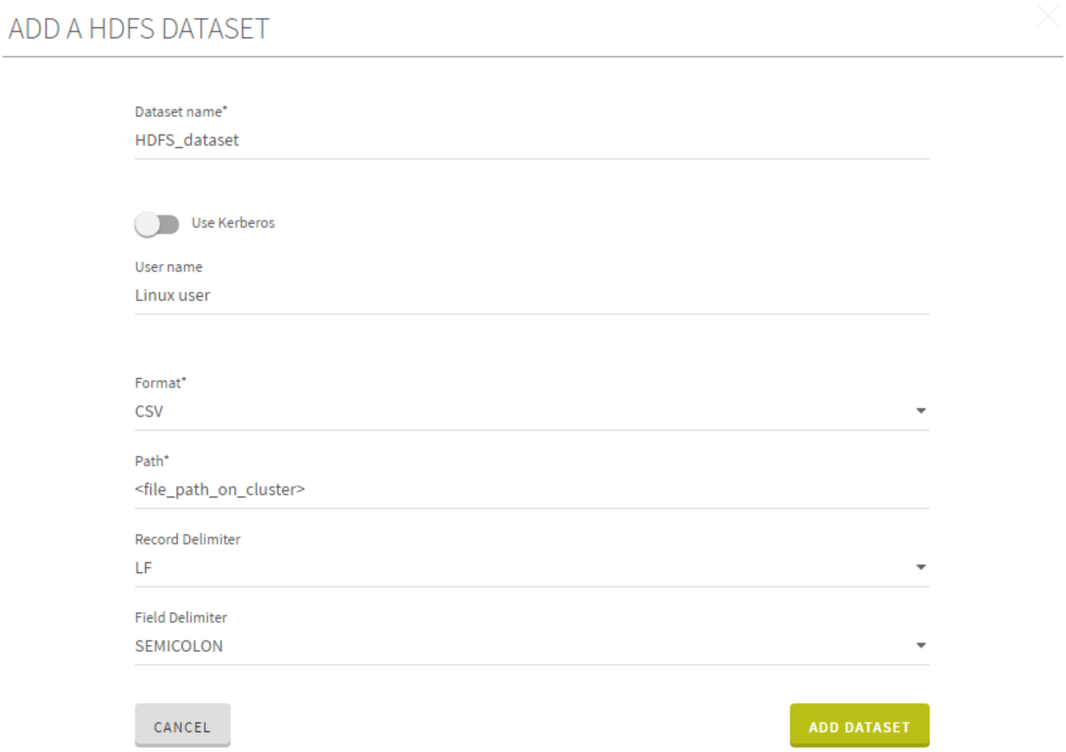
Importing data from the cluster

You will access your 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 a 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>

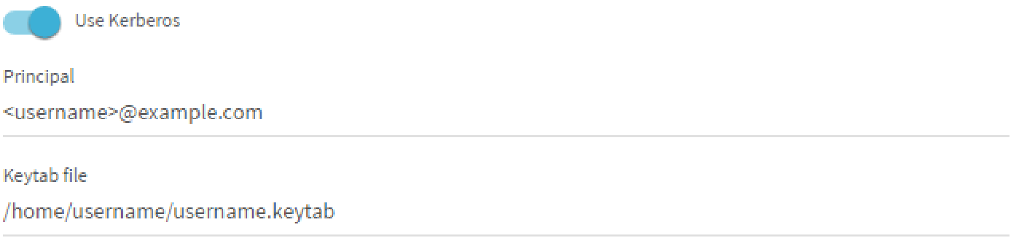
Record Delimiter
LF

Field Delimiter
SEMICOLON

3. In the **Dataset name** field, enter the name you want to give your dataset., `HDFS_dataset` in this example.
4. In the **User name** field enter the name of the Linux user on the cluster.

   This user must have the reading rights on the file that you want to import.
5. For this example, leave the **Use Kerberos** check box unselected.

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

   Use Kerberos

   Principal
   <username>@example.com

   Keytab file
   /home/username/username.keytab

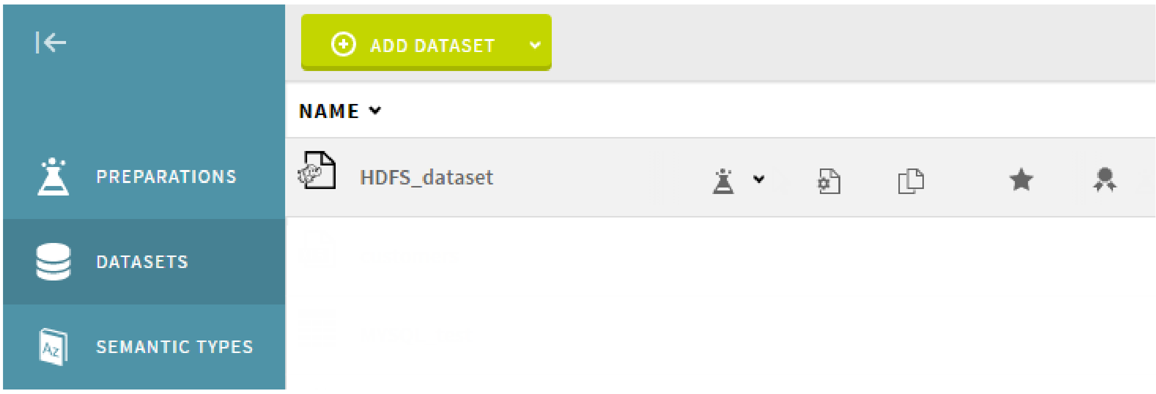
   The `keytab` file must be accessible by the Spark Job Server.

   You can manually configure Talend Data Preparation to display a default value in those fields.
6. In the **Format** field, select the format in which your data was stored in the cluster, `.csv` in this case.
7. In the **Path** field, enter the complete URL of your file in the Hadoop cluster.
8. Click **Add Dataset**.

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.

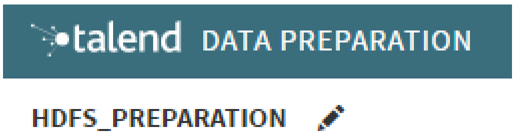
Your dataset is now available in the **Datasets** view of the application home page.



Saving your preparation

Now that you have successfully created a dataset with the data coming from the cluster, you will save it with a proper name.

1. Click the grey arrow on the left side of the screen to expand the side panel.
2. In the preparation name field, enter `HDFS_Preparation`.
3. Press **Enter** to confirm the operation.



You have named your preparation, and from now on, every change you make to your preparation will automatically be saved.

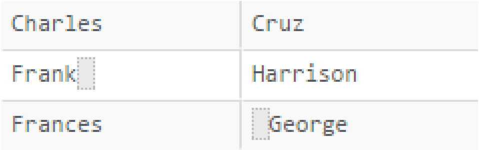
Cleansing your data

Now that your preparation has been saved, you can start working on the customer data, like with any other dataset, and choose among all the usual functions.

The dataset that you have imported originally contains 20,000 rows but only a sample of the first 10,000 by default rows is displayed. Don't worry, all the preparation steps that you add can be applied to the whole dataset.

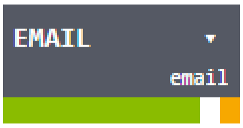
You will perform some basic cleansing operations, to ensure that all the data contained in the dataset is valid and free of errors.

You can for example notice the presence of unnecessary whitespaces in some entries of the **First_Name** and **Last_Name** columns.



Charles	Cruz
Frank	Harrison
Frances	George

The quality bar under each column also indicates that your data contains rows with empty or invalid cells. The **Email** column, for example, contains both.

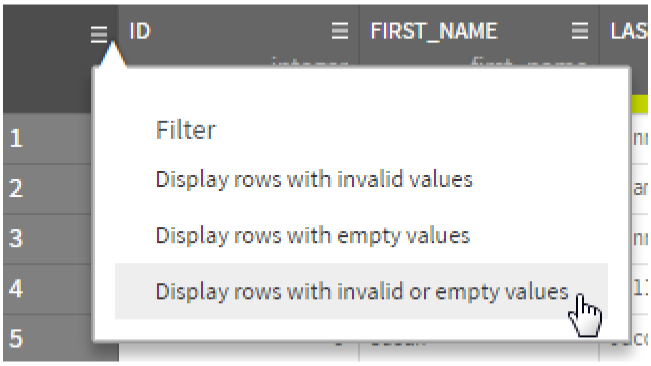


You are going to delete all the empty and invalid rows from the preparation in a single action, and remove the formatting errors in the columns containing the customer names.

1. Click the header of the **First_Name** column.
2. While keeping the **Ctrl** button pressed, click the header of the **Last_Name** column.

   The two columns are now selected, and you can apply a function to both columns in one action.
3. In the **Functions panel**, search for the **Remove trailing and leading characters** function and click it to open the options panel.
4. In the **Padding character** drop-down list, select **whitespace** and click **Submit**.

   Blank spaces have been removed from the selected columns.
5. Click the white arrow on the top left of the grid and select **Display rows with invalid or empty values**.

   
   ID
   FIRST_NAME
   LAS
   1
   2
   3
   4
   5
   Filter
   Display rows with invalid values
   Display rows with empty values
   Display rows with invalid or empty values

   A filter has been applied on your data, and only the rows with empty or invalid cells are displayed, making it easier for you to delete them in one go.
6. In the **Functions panel**, click **Delete these Filtered Rows** to apply the corresponding function.

   All the filtered lines have been deleted, and you can now clear the filter by clicking the garbage bin icon in the filter bar.

In two simple actions, you have removed all the errors contained in your dataset and improved the quality of your data.

The quality bar for each column is now completely green, indicating that there is no invalid data left in your preparation.

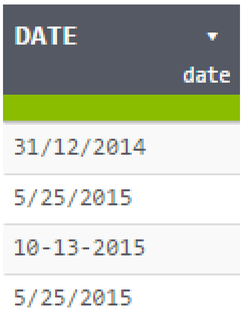
Harmonizing the date format

The **Date** column stores the information about the subscription date to the online service.

In this case, you have noticed that at least three different types of file format are used in the same column:

- `dd/MM/yyyy`

- `MM/dd/yyyy`
- `MM-dd-yyyy`



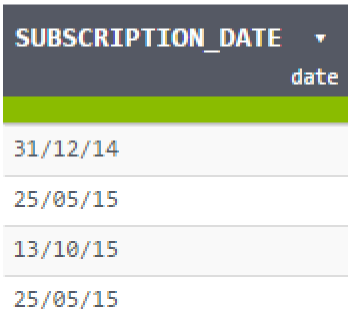
DATE ▾ date
31/12/2014
5/25/2015
10-13-2015
5/25/2015

You will give the column a more meaningful name, and harmonize the date format used in the column.

1. To rename the **Date** column, double click the column name.
2. Type `subscription_date` and press **Enter**.
   The column has been renamed to better describe its content, that you will now harmonize.
3. In the **Functions panel**, search for the **Change Date Format** function and click it.

   A menu with the options for the dedicated function opens.
4. In the **New Format** list, select the new format to apply to the whole column, in this case, click **French standard** to use the `dd/MM/yy` format.

The column containing the subscription dates is now easier to read, with a proper title and a unified format.



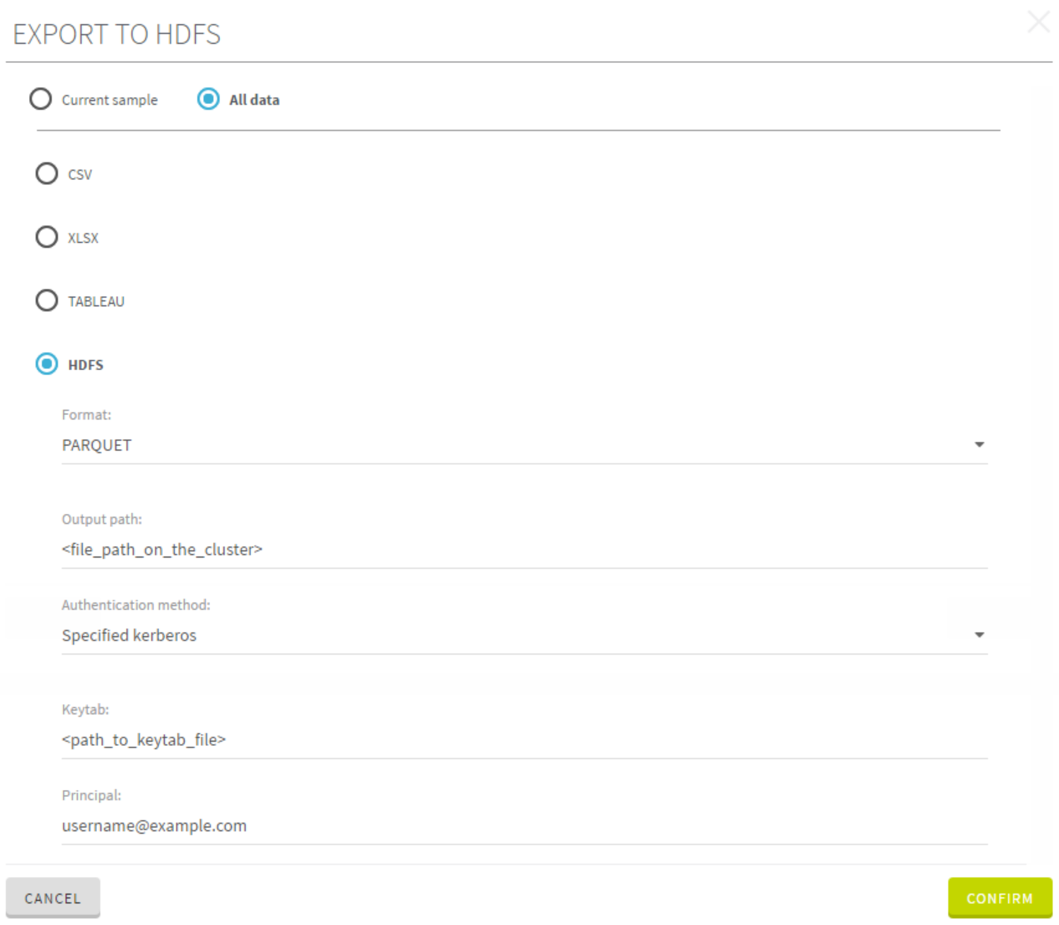
SUBSCRIPTION_DATE ▾ date
31/12/14
25/05/15
13/10/15
25/05/15

Exporting your preparation to the cluster

Now that your are done preparing your data, you will export it back to the cluster, but as a `Parquet` file this time.

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: PARQUET

Output path: <file_path_on_the_cluster>

Authentication method: Specified kerberos

Keytab: <path_to_keytab_file>

Principal: username@example.com

2. Select the **All data** radio button so that the whole data is prepared, and not just the sample you worked on.
3. Select the **HDFS file** radio button to export your data to the Hadoop cluster.

   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.
4. Select the **Parquet** format.
5. In the **Output path** field, enter the complete URL to your prefered location on the cluster to save the exported file.

   You can manually configure Talend Data Preparation to display a default value in the **Output Path** field.
6. Select **Specified kerberos** as authentication method.
7. Specify your principal and the path to your `keytab` file.

   If you choose **Default Kerberos**, the values for the `keytab` file path and the principal will be the ones entered in Talend Data Preparation configuration file.

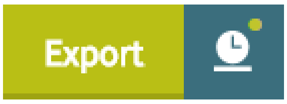
   In any case, the path must point to a `keytab` file that is accessible to all the workers on the cluster.

   Select the **Simple** authentication if you are not using Kerberos.
8. Click **Confirm**

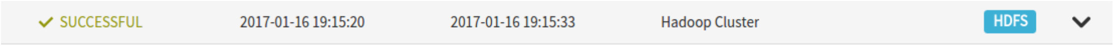
   You export starts in the background, and is now being processed directly on the cluster.

   Note that if a preparation contains actions that only affect a single row, or cells, they will be skipped during the export process. A warning will be displayed before the export if your preparation contains such actions.

9. Click the **Export history** button in the application header bar to check the status of the export.



Among other information, you can see that the export was successful.



✓ SUCCESSFUL 2017-01-16 19:15:20 2017-01-16 19:15:33 Hadoop Cluster HDFS

Your data has been processed and saved as a `parquet` file, without leaving the cluster.