Talend Data Preparation
Free Desktop - How-tos
Contents

Concepts and Principles............................................................................................................................. 4
- What is a dataset?........................................................................................................................................ 4
- What is a function?................................................................................................................................... 4
- What is a preparation?............................................................................................................................... 4
- What is a recipe?....................................................................................................................................... 4
- Using charts............................................................................................................................................. 5
  - Chart displayed for date data................................................................................................................ 5
  - Chart displayed for string data.............................................................................................................. 5
  - Charts displayed for numerical data...................................................................................................... 6
  - Vertical bar chart.................................................................................................................................. 6
  - Range slider.......................................................................................................................................... 7
  - Box Plot................................................................................................................................................ 8
  - Horizontal bar chart.............................................................................................................................. 9

Preparing client data to upload it to a marketing solution.................................................. 10
- Removing empty records from a column.............................................................................................. 10
- Removing invalid records from a column.............................................................................................. 11
- Extracting email address parts.............................................................................................................. 12
- Removing unnecessary blank spaces in text records......................................................................... 13
- Saving a preparation............................................................................................................................... 14
- Exporting the results of your recipe....................................................................................................... 14

Cleansing data coming from a human resource management system.............. 15
- Changing the date format....................................................................................................................... 15
- Extracting the bank account number.................................................................................................. 16
- Saving a preparation............................................................................................................................... 17
- Exporting the results of your recipe....................................................................................................... 17

Recreating email addresses before uploading them to a marketing
solution....................................................................................................................................................... 18
- Removing empty records from a column.............................................................................................. 18
- Removing unnecessary blank spaces in text records.......................................................................... 19
- Duplicating a column.............................................................................................................................. 20
- Renaming a column............................................................................................................................... 21
- Changing the case to lower case........................................................................................................... 22
- Dynamically using the data from another dataset............................................................................. 23
- Merging the content of two or more columns.................................................................................... 25
- Deleting a column.................................................................................................................................. 26
- Saving a preparation............................................................................................................................... 27
- Exporting the results of your recipe....................................................................................................... 28

Using "if" conditions: consolidating a list of phone numbers coming from
a CRM solution......................................................................................................................................... 28
- Creating a column to receive the consolidated numbers..................................................................... 28
- Duplicating the MOBILE column........................................................................................................ 29
Renaming the **MOBILE_COPY** column...................................................................................... 29
Filling empty mobile phone numbers with landline numbers...................................................... 30
Identifying if there are mobile phone numbers among the landline numbers............................ 31
Replacing the consolidated phone numbers with the mobile phone numbers identified in the previous step................................................................................................................................... 32
  Create a filter on the mobile phone numbers....................................................................... 32
  Replacing the consolidated phone numbers with the mobile phone numbers only for the filtered values............................................................................................................................... 33
Deleting the **PHONE MATCHING** column........................................................................... 34
Removing empty records from a column..................................................................................... 35
Saving a preparation.................................................................................................................. 36
Exporting the results of your recipe............................................................................................ 37
Concepts and Principles

Here are the concepts and principles that will help you understand Talend Data Preparation Free Desktop.

What is a dataset?

A dataset holds the raw data that can be used as the raw material for one or more recipes. It is presented as a table on which you can apply recipes without affecting the original data. As they are not altered by the recipes, datasets can be reused across preparations.

What is a function?

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

What is a 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, although not in the free version). A preparation takes one dataset and applies a recipe to produce an outcome. The original dataset is never modified.

What is a 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 Free Desktop, 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.

Do not look for a save button: every update of the recipe is automatically saved in the preparation all the time.

Using charts
Charts are a graphical representation of the distribution of the values contained in the selected column. The charts displayed are different according to the type of data contained in the selected column. There are three different combinations:

- the charts displayed for numerical data,
- the chart displayed for date data, and
- the charts displayed for string data.

Chart displayed for date data
When the type of data contained in the selected column is a date, the chart for date data is displayed.

This chart is a **vertical bar chart**.

Chart displayed for string data
When the type of data contained in the selected column is a string, the chart for string data is displayed.
This chart is a horizontal bar chart.

**Charts displayed for numerical data**

When the type of data contained in the selected column is an integer or a decimal, charts for numerical data are displayed.

These charts are:

- the **vertical bar chart**,  
- the **range slider**, and  
- the **box plot**.

**Vertical bar chart**

The vertical bar chart is a histogram displayed in the **CHART** tab when the selected column contains numerical or date data.

![Histogram chart](image)

This histogram represents the distribution of the data.

When it used with date data, the dates are displayed in the X axis. When it is used with numerical data, the values are displayed using the range slider.

It is an interactive chart: you can create a new filter by clicking one of the bars of the chart.

Also, if you point your mouse over one of the bars, the number of occurrences for the selected range is displayed alongside the endpoints of the interval.
Intervals are written as right half-open intervals from the left value (included) to the right value (excluded). For example, if the interval is [1950, 1960[, it means that it includes all the numbers from 1950 to 1959.

**Range slider**

The range slider is displayed under the vertical bar chart in the **CHART** tab when the selected column contains numerical or date data.

The range slider represents the extent of numerical column.

It allows you to create a filter that reduces the range of data displayed in the selected column.

You can also use the **Min** and **Max** fields to precisely define the range of data. This can be useful to filter values which are not present in the current dataset but that could be present in another dataset to which the current preparation will also apply.
The value you enter in the **Min** and **Max** fields should be set between \(-9,007,199,254,740,991\) and \(9,007,199,254,740,991\) and the decimal separator to be used is the period (\(\cdot\)).

When a filter is defined using the range slider, the color of the vertical bar chart changes to reflect the filter selection.

---

**Related concepts**

- **Vertical bar chart** on page 6
  
  The vertical bar chart is a histogram displayed in the **CHART** tab when the selected column contains numerical or date data.

- **Box Plot**
  
  The box plot is displayed in the **ADVANCED** tab when the selected column contains numerical data.
It allows you to have a graphical summary of your numerical data by showing its mean, its median, the quartiles, the maximum value and the minimum one.

**Related concepts**

- **Using charts** on page 5
  Charts are a graphical representation of the distribution of the values contained in the selected column.

- **Vertical bar chart** on page 6
  The vertical bar chart is a histogram displayed in the CHART tab when the selected column contains numerical or date data.

**Horizontal bar chart**

The horizontal bar chart is a histogram displayed when the selected column contains string data.
This histogram represents the distribution of the data of the 15 more represented occurrences.

It is interactive: you can create a new filter by clicking one of the bars of the chart.

Also, if you point your mouse over one of the bars, the number of occurrences for the selected string is displayed.

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 this dataset in order to be able to upload it to a marketing solution.

Removing empty records from a column

The white part of the Quality Bar indicates that the column contains empty records. You may want to remove the lines containing these empty records.

To remove the lines containing empty records for a particular column, proceed as follows:

1. Click the white part of the Quality Bar and click Select empty records for to preview the lines containing the empty records. You can filter several columns using the Select empty records for function in order to create more complex filters.
2. Click **Delete the Lines with Empty Cell** to delete the lines that do not contain any data. You can also click this option in the **Functions panel** on the right-hand side of the application.

3. Repeat these actions on other columns to remove lines containing empty cells.

The lines containing empty records are removed from the dataset.

**Removing invalid records from a column**

The orange part of the **Quality Bar** indicates that the column contains invalid records. You may want to remove the lines containing these invalid records.

To remove the lines containing invalid records, proceed as follows:

1. Click the orange part of the **Quality Bar** and click **Select invalid records for** to preview the lines containing the invalid records.
2. Click **Delete the Lines with Invalid Cell** to delete the lines that contain invalid data. You can also click this option in the **Functions panel** on the right-hand side of the application.

3. Repeat these actions on other columns to remove lines containing invalid cells.

The lines containing invalid records are removed from the dataset.

**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). These parts can be extracted and copied to two new columns.

To extract the local part and the domain part from email addresses, proceed as follows:

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.
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.

To remove unnecessary blank spaces for a particular column, proceed as follows:

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

2. In the **Functions panel**, type **Remove Whitespaces (Trailing and Leading)** and click the result to execute the associated function.
3. Repeat this action for every column containing blank spaces.

Blank spaces are removed from the selected column.

Saving a preparation

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

To save a preparation, proceed as follows:

1. On the left part of **Talend Data Preparation Free Desktop**, enter a name for your preparation.
2. Click the green tick icon to validate your choice.

Your preparation is saved in **Talend Data Preparation Free Desktop**.

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

Exporting the results of your recipe

Once your recipe is complete, you may want to export the sample dataset you have cleansed using your preparation.

To export the results of your recipe, proceed as follows:

1. Click the arrow next to 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**, 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 bank account number from the IBAN number for the French accounts.

**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.

To change the date format of a column containing dates, proceed as follows:

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 **other**.

![SUGGESTION Change Date Format...](image)

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.

   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.

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

To extract part of a text from a column, proceed as follows:

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 From index in the From list and type 17.
   This will start the selection at the 17th character.

4. Choose To N before end in the To list and type 2.
   This will end the selection at the 3rd character before the end.

5. 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.

Saving a preparation

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

To save a preparation, proceed as follows:

1. On the left part of **Talend Data Preparation Free Desktop**, enter a name for your preparation.
2. Click the green tick icon to validate your choice.

Your preparation is saved in **Talend Data Preparation Free Desktop**.

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

Exporting the results of your recipe

Once your recipe is complete, you may want to export the sample dataset you have cleansed using your preparation.
To export the results of your recipe, proceed as follows:

1. Click the arrow next to 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, 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 there are no email addresses. 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.

**Removing empty records from a column**

The white part of the Quality Bar indicates that the column contains empty records. You may want to remove the lines containing these empty records.

To remove the lines containing empty records for a particular column, proceed as follows:

1. Click the white part of the Quality Bar and click Select empty records for to preview the lines containing the empty records. You can filter several columns using the Select empty records for function in order to create more complex filters.
2. Click **Delete the Lines with Empty Cell** to delete the lines that do not contain any data. You can also click this option in the **Functions panel** on the right-hand side of the application.

3. Repeat these actions on other columns to remove lines containing empty cells.

The lines containing empty records are removed from the dataset.

### 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.

To remove unnecessary blank spaces for a particular column, proceed as follows:

1. Select a column containing text with blank spaces or a column that you think may contain some.
2. In the **Functions panel**, type **Remove Whitespaces (Trailing and Leading)** and click the result to execute the associated function.

3. Repeat this action for every column containing blank spaces.

Blank spaces are removed from the selected column.

**Duplicating a column**

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

To duplicate a column, proceed as follows:

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**.
4. For this example, repeat these actions for the **LAST_NAME** column.

The selected column is duplicated.

**Renaming a column**

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

To rename a column, proceed as follows:

1. Click the column you want to rename.
2. Perform one of the following actions:
   - Double-click the name of the column, or
   - Use the **Rename Column** option in the column contextual menu.
3. Enter a new name for the selected column.

   In this example, you can rename the **FIRST_NAME_COPY** column as **EMAIL_FIRST_NAME** and the **LAST_NAME_COPY** column as **EMAIL_LAST_NAME**.

   The selected column is 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.

To change the case to lower case for the text contained in a particular column, proceed as follows:

1. Select a 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 this action for every column you want to modify.

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

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.

To dynamically use the data from another dataset, proceed as follows:

1. Select the column on which you want to perform the lookup, the **COMPANY** column in our example. This will be the source column for your data.

2. Click the 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. This dataset contains the target column for your data.
4. Select the Add to Lookup check box under every column you want to include in your lookup, **EMAIL_DOMAIN** in this example.

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.

**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.

To merge the content of columns, proceed as follows:

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.
   
   In this example, use the **EMAIL_FIRST_NAME** column.

2. In the **Functions panel**, type **Concatenate with...** and click the result to display the options of the associated function.

3. In the **Column** list, choose **EMAIL_LAST_NAME**.

4. In the **Separator** field, enter . to add a period as a separator.

5. Click the **SUBMIT** button to apply the function.
6. Proceed the same way to merge the column you just created with the `EMAIL_DOMAIN` one, using the @ symbol as a separator.

The content of two columns is merged using a concatenation.

Deleting a column

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

To delete a column, proceed as follows:

1. Click the column you want to delete, `EMAIL_FIRST_NAME` in this example.
2. Click the white arrow or right-click the column to open the contextual menu.
3. Click `Delete Column`. 
In this example, you can proceed the same way to delete the EMAIL_LAST_NAME column and the EMAIL_DOMAIN one.

The selected column is deleted.

Saving a preparation

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

To save a preparation, proceed as follows:

1. On the left part of Talend Data Preparation Free Desktop, enter a name for your preparation.
2. Click the green tick icon to validate your choice.

Your preparation is saved in Talend Data Preparation Free Desktop.

Every modification you make from now on will automatically be saved in the preparation you created.
Exporting the results of your recipe

Once your recipe is complete, you may want to export the sample dataset you have cleansed using your preparation.

To export the results of your recipe, proceed as follows:

1. Click the arrow next to 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**, choose a name for the file to export.

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

Using "if" conditions: 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 Free Desktop**, you will create filters.

Creating a column to receive the consolidated numbers

Before working on the data, you will create a new column to receive the consolidated numbers.
In order to do that, you will:

1. Duplicate the MOBILE column in order to have a copy of the data.
2. Rename the new column in order to give it a meaningful name.

**Duplicating the MOBILE 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 the column.

To duplicate the MOBILE column, proceed as follows:

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.

The MOBILE column is duplicated.

**Renaming the MOBILE_COPY column**

In order to better identify the new column, you will give it a meaningful name.

To rename the MOBILE_COPY column, proceed as follows:
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, or
   - Use the Rename Column option in the contextual menu for the column.

3. Enter a new name for the selected column.
   In this example, rename the MOBILE_COPY column as PHONE NUMBER TO USE.

The MOBILE_COPY column is renamed.

Filling empty mobile phone numbers with landline numbers

The white part of the Quality Bar indicates that the column contains empty records. You will fill these records with landline numbers.

To fill the lines containing empty mobile phone numbers with landline numbers, proceed as follows:

1. Select a column containing empty records, PHONE NUMBER TO USE in this example.
   You can identify columns containing empty records using the Quality Bar. The white part of it indicates that the column contains empty records.
2. In the Functions panel, type Fill Empty Cells with Text... and click the result to open the options for the associated function.

Fill in the options as shown in the following image:

![Fill Empty Cells with Text](image)

3. Click the SUBMIT button to apply the function.

This fills all the empty cells in the MOBILE column with data from the PHONE column.

Identifying if there are mobile phone numbers among the landline numbers

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

To identify if there are mobile phone numbers among the landline numbers, proceed as follows:

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.

   Fill in the options as shown in following image:

   ![Matches Pattern dialog]

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.

Replacing the consolidated phone numbers with the mobile phone numbers identified in the previous step

   Now that the mobile phone numbers that are present among the landline numbers are identified, you want to copy them to the PHONE NUMBER TO USE column.

   In order to do that, you will:

   1. Create a filter that selects the identified mobile phone numbers.

   2. Replace the consolidated phone numbers with the mobile phone numbers in these filtered values.

Create a filter on the mobile phone numbers

   In order to replace only the line that matches the filter defined in the previous step, you will create a filter that selects the identified mobile phone numbers.

   To create a filter, proceed as follows:

   1. In the Search and Filter field, enter some data you want to use as a filter, true in this example.
2. Choose the column on which you want the filter to apply, PHONE_MATCHING in this example.

The data is filtered on the value true.

Replacing the consolidated phone numbers with the mobile phone numbers only for the filtered values

Now that only the mobile phone numbers are selected, you will replace the consolidated phone numbers.

To do this, proceed as follows:

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.

Fill in the options as shown in the following image:
3. Make sure the **Filtered Lines** 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, **MOBILE** in this example, is replaced with the data from the **PHONE** column.

**Deleting the PHONE MATCHING column**

As the **PHONE MATCHING** column is not used anymore, you can delete it.

To delete the **PHONE MATCHING** column, proceed as follows:

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 the column contains empty records. You may want to remove the lines containing these empty records.

To remove the lines containing empty records for a particular column, proceed as follows:

1. Click the white part of the Quality Bar and click Select empty records for to preview the lines containing the empty records. You can filter several columns using the Select empty records for function in order to create more complex filters.
2. Click **Delete the Lines with Empty Cell** to delete the lines that do not contain any data. You can also click this option in the **Functions panel** on the right-hand side of the application.

3. Repeat these actions on other columns to remove lines containing empty cells.

The lines containing empty records are removed from the dataset.

**Saving a preparation**

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

To save a preparation, proceed as follows:

1. On the left part of **Talend Data Preparation Free Desktop**, enter a name for your preparation.
2. Click the green tick icon to validate your choice.
Your preparation is saved in **Talend Data Preparation Free Desktop**.

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

**Exporting the results of your recipe**

Once your recipe is complete, you may want to export the sample dataset you have cleansed using your preparation.

To export the results of your recipe, proceed as follows:

1. Click the arrow next to 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**, choose a name for the file to export.

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