Data Manipulation with Talend Data Preparation Free Desktop

พิชิตชัย พิมพ์โคตร

DPU

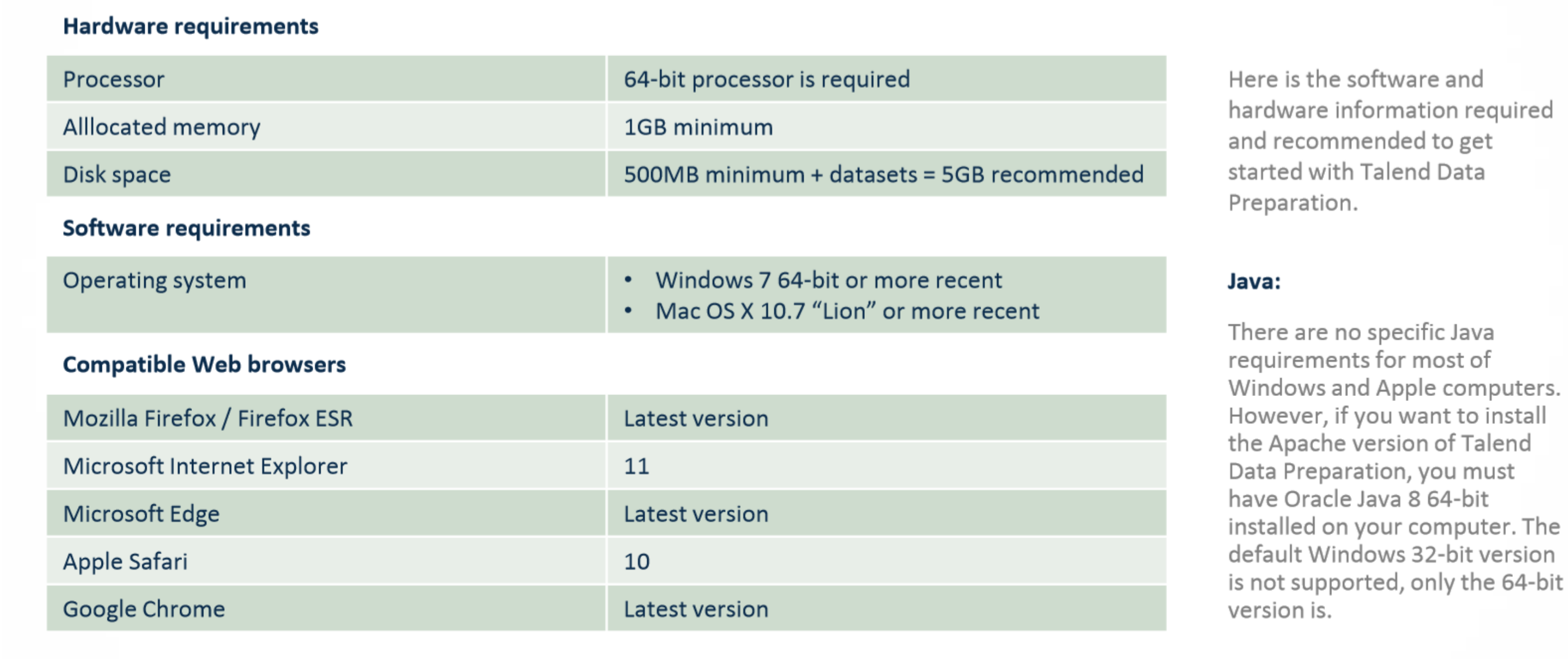
CITE

College of Innovative Technology and Engineering Dhurakij Pundit University

Table of contents

Page

Requirements for Talend Data Preparation Free Desktop



Hardware requirements

Processor	64-bit processor is required
Alllocated memory	1GB minimum
Disk space	500MB minimum + datasets = 5GB recommended

Software requirements

Operating system	Windows 7 64-bit or more recent Mac OS X 10.7 “Lion” or more recent

Compatible Web browsers

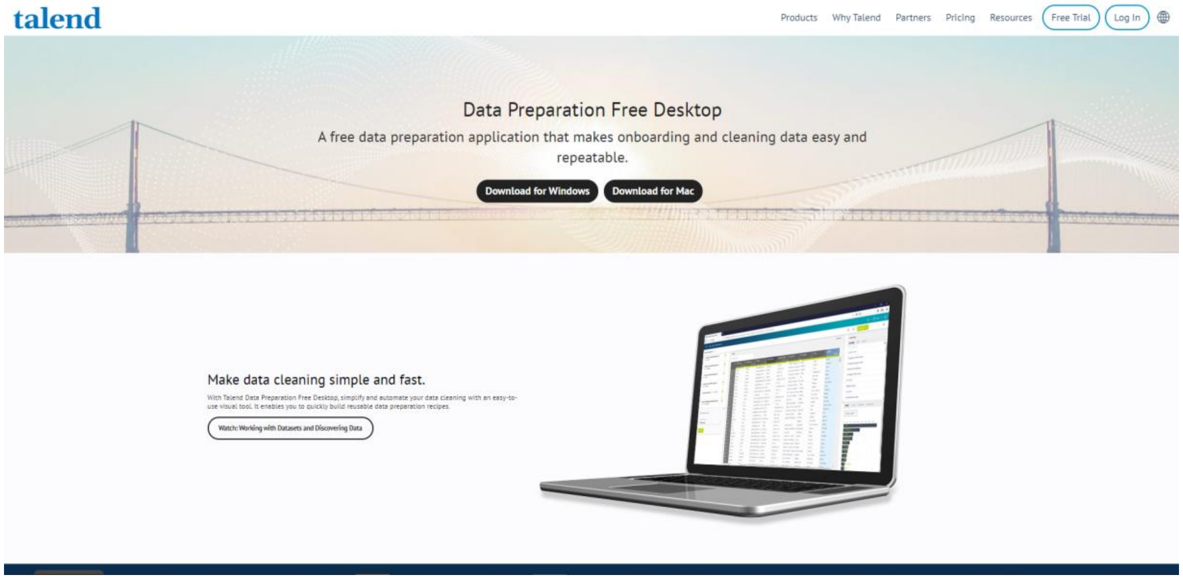
Mozilla Firefox / Firefox ESR	Latest version
Microsoft Internet Explorer	11
Microsoft Edge	Latest version
Apple Safari	10
Google Chrome	Latest version

Here is the software and hardware information required and recommended to get started with Talend Data Preparation.

Java:

There are no specific Java requirements for most of Windows and Apple computers. However, if you want to install the Apache version of Talend Data Preparation, you must have Oracle Java 8 64-bit installed on your computer. The default Windows 32-bit version is not supported, only the 64-bit version is.

Download Installer Package



talend

Products Why Talend Partners Pricing Resources Free Trial Log In

Data Preparation Free Desktop

A free data preparation application that makes onboarding and cleaning data easy and repeatable.

Download for Windows Download for Mac

Make data cleaning simple and fast.

With Talend Data Preparation Free Desktop, simplify and automate your data cleaning with an easy-to-use visual tool. It enables you to quickly build reusable data preparation recipes.

Watch: Working with Datasets and Discovering Data

Download Talend Data Preparation here
>> Click <<

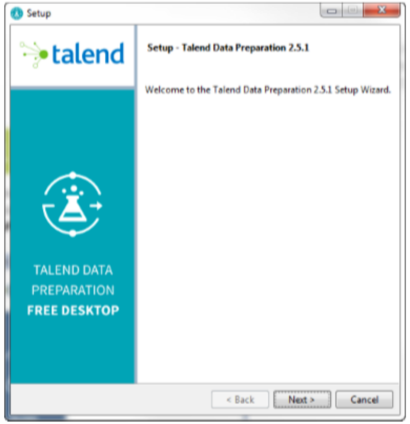
Select your operating system, and the download starts automatically.

Link download : https://www.talend.com/products/data-preparation/data-preparation-free-desktop/?qt-product_tos_download_new=5

Set-up Talend Data Preparation on Windows

1

Locate the file you have just downloaded and double click **Talend-DataPreparation-Free-Desktop-2.5.exe**


Setup

talend

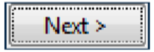
Setup - Talend Data Preparation 2.5.1

Welcome to the Talend Data Preparation 2.5.1 Setup Wizard.

TALEND DATA PREPARATION FREE DESKTOP

2

Click **Next** through the setup and use the default settings.

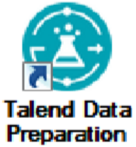


3

Click **Finish** once the Install is complete.

4

Use the Desktop icon or the shortcut on the Start menu to begin using the Talend Data Preparation tool.



talend

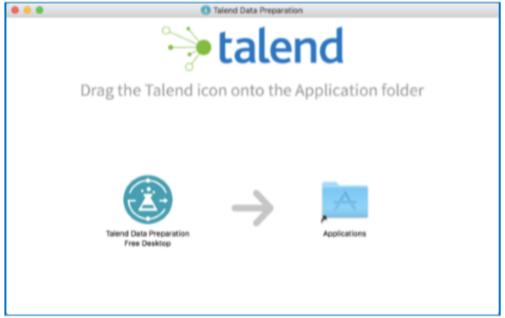
Set-up Talend Data Preparation on Mac

1

Double-click the Talend-DataPreparation-Free-Desktop-2.5.dmg file to open the package.

2

Drag and drop into the Applications folder.



Drag the Talend icon onto the Application folder

3

Talend Data Preparation will now be in your list of **Applications**. Locate the icon and double click to open the application.

4

To disable **App Nap** and ensure optimal performance, follow this quick procedure:

1. Open the Terminal from the `/Applications/Utilities` folder.

2. Enter the following command:

```
defaults write
org.talend.dataprep
NSAppSleepDisabled
-bool YES
```

talend

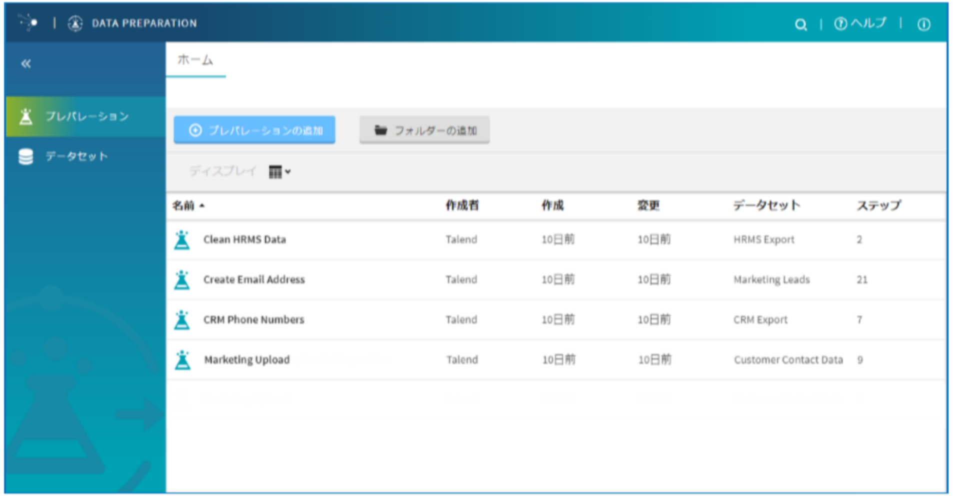
Configuring the language of the interface

1. Open the `<TDP_Installation_Path>/dataprep/config/application.properties` configuration file.

2. For the `dataprep.locale` parameter, enter one of the three supported values:

   - **en-US** for English
   - **fr-FR** for French
   - **ja-JP** for Japanese

3. Restart Talend Data Preparation Free Desktop.

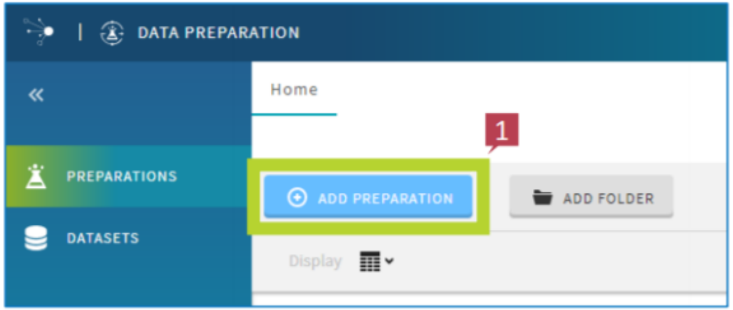


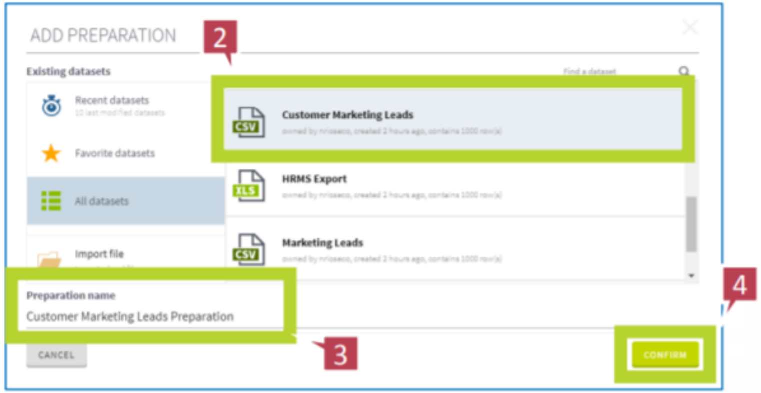
名前 ▴	作成者	作成	変更	データセット	ステップ
Clean HRMS Data	Talend	10日前	10日前	HRMS Export	2
Create Email Address	Talend	10日前	10日前	Marketing Leads	21
CRM Phone Numbers	Talend	10日前	10日前	CRM Export	7
Marketing Upload	Talend	10日前	10日前	Customer Contact Data	9

How to add a preparation?

To get started on the example:

1. Click the **Add Preparation** button from the **Preparations** view.
2. The **Add Preparation** dialog opens. Click the **Customers Marketing Leads** dataset from the **All Datasets** list.
3. Choose a **name** for your preparation.
4. Click **Confirm** to open the preparation and start cleansing the data from the dataset.





Preparation name
Customer Marketing Leads Preparation

Manipulate data

Simple cleansing examples

First, let's fix the **Name** column:

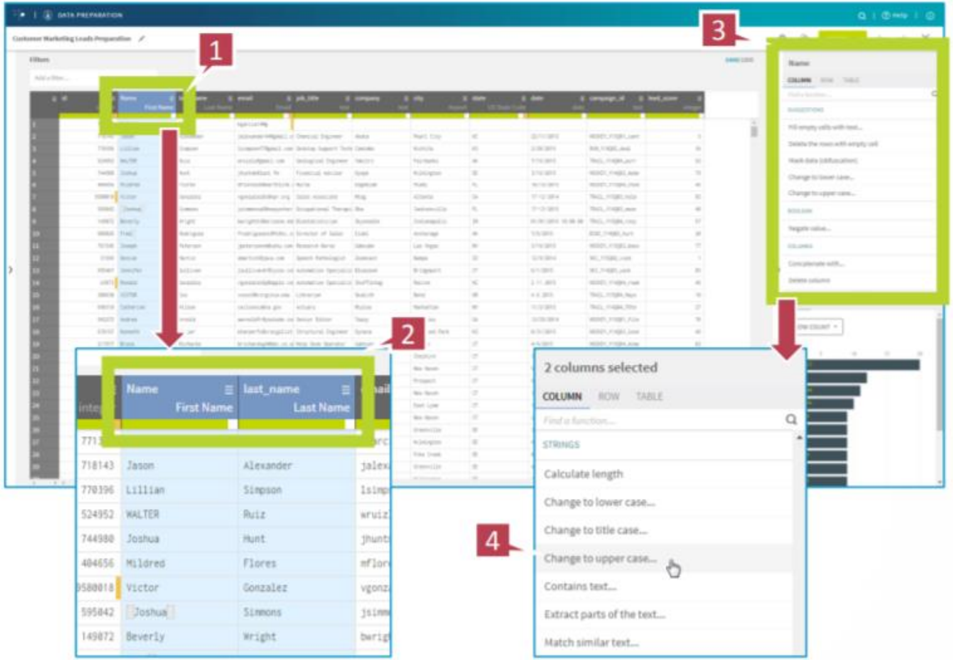
1. Click the header of the **Name** column.

2. While pressing the **Ctrl** key, click the header of the **last_name** column. Both columns are now selected. You can also use **Shift + click** to select multiple columns.

3. In the top right corner is a box with all the **Functions** available. You can search for functions or use one of the suggested functions to improve your data.

4. You may need to scroll, down depending on your screen resolution, to find the **Change to upper case** function. **Hover over the function to preview its effect on the data. Click the function to apply the changes to the two selected columns.**

Here we are cleaning up the customer name fields to do some basic standardization, You can see that there are mixed case names, leading and training Spaces and the last name has been defined as an incorrect type.



	Name First Name	last_name Last Name
718143	Jason	Alexander
770396	Lillian	Simpson
524952	WALTER	Ruiz
744980	Joshua	Hunt
404656	Mildred	Flores
9580018	Victor	Gonzalez
595042	Joshua	Simmons
149072	Beverly	Wright

Manipulate data

Simple cleansing examples

To execute basic formatting and cleansing:

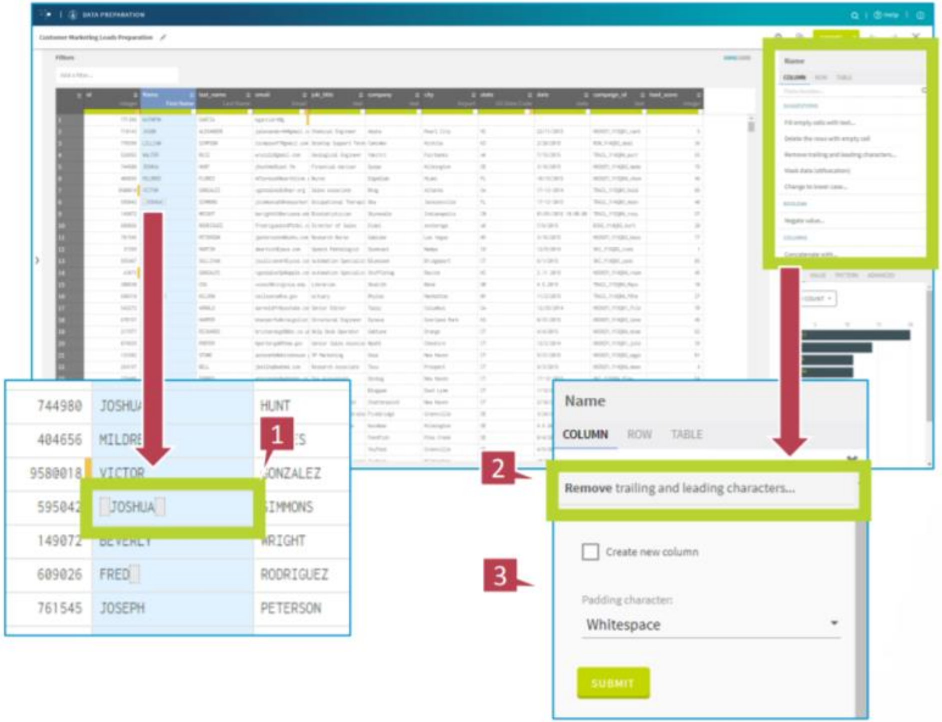
Name column, continued.

1. While looking at the data, you will see grey boxes in front or behind some names, for example “Joshua”.

2. To remove those grey boxes, search for and select the **Remove trailing and leading characters** function.

3. Leave the **Create new column** checkbox clear. In the **Padding character** drop-down list, select **Whitespace** and click **Submit**.

Several functions, including this one, allow you to output the result of the transformation in a new column by selecting the **Create new column** checkbox. If you do not select it, the function will be applied in the current column.



744980	JOSHUA	HUNT
404656	MILDRE	S
9580018	VICTOR	GONZALEZ
595042	JOSHUA	SIMMONS
149072	BEVERLY	WRIGHT
609026	FRED	RODRIGUEZ
761545	JOSEPH	PETERSON

Name

COLUMN ROW TABLE

Remove trailing and leading characters...

□ Create new column

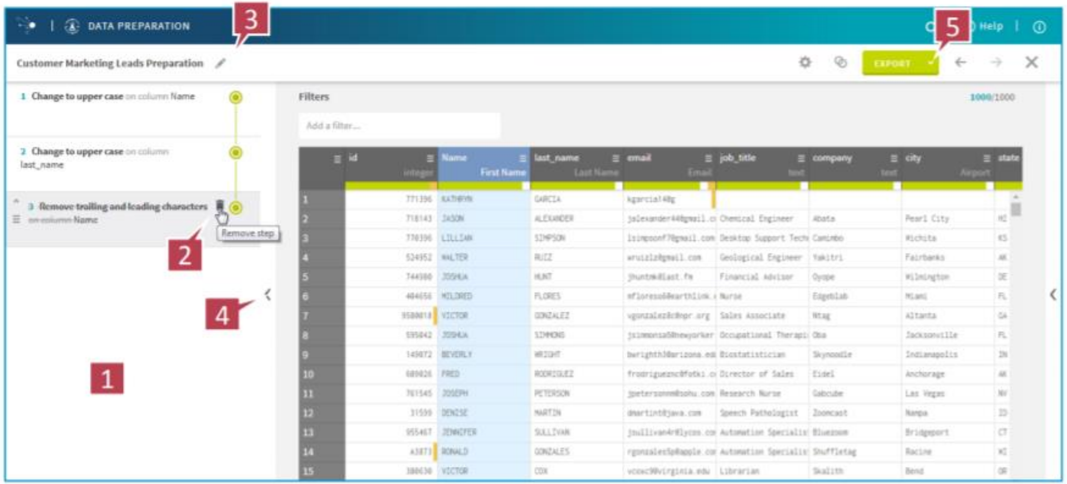
Padding character:
Whitespace

SUBMIT

Manipulate data

Recipes

1. Every time you apply a function, it is added to the recipe panel on the left.
2. **To delete a recipe item**, point your mouse over the line item and click the trash can icon.
3. **To rename a preparation**, click the pencil icon and enter a new name.
4. The recipe panel can be hidden by clicking the arrow.
5. To export the result of your preparation, click **Export** and select a file type.



Customer Marketing Leads Preparation

1 **Change to upper case** on column Name

2 **Change to upper case** on column last_name

3 **Remove trailing and leading characters** on column Name

Filters

Add a filter...

1000/1000

	id	Name	last_name	email	job_title	company	city	state
	integer	First Name	Last Name	Email	text	text	Airport	
1	771396	KATHRYN	GARCIA	kgarcia14@g				
2	718143	JASON	ALEXANDER	jalexander44@gmail.c	Chemical Engineer	Abata	Pearl City	HI
3	770396	LILLIAN	SIMPSON	lsimpsonf7@gmail.com	Desktop Support Tech	Camimbo	Wichita	KS
4	524952	WALTER	RUIZ	wruiz1z@gmail.com	Geological Engineer	Yakitri	Fairbanks	AK
5	744980	JOSHUA	HUNT	jhuntm@last.fm	Financial Advisor	Oyope	Wilmington	DE
6	404656	MILDRED	FLORES	mfloreso6@earthlink.	Nurse	Edgeblab	Miami	FL
7	9580018	VICTOR	GONZALEZ	vgonzalez8c@npr.org	Sales Associate	Ntag	Altanta	GA
8	595042	JOSHUA	SIMMONS	jsimmonsa5@newyorker	Occupational Therapi	Oba	Jacksonville	FL
9	149072	BEVERLY	WRIGHT	bwrighth3@arizona.ed	Biostatistician	Skynoodle	Indianapolis	IN
10	680026	FRED	RODRIGUEZ	frodrigueznc@fotki.c	Director of Sales	Eidel	Anchorage	AK
11	761545	JOSEPH	PETERSON	jpetersonnm@sohu.com	Research Nurse	Gabcube	Las Vegas	NV
12	31599	DENISE	MARTIN	dmartint@java.com	Speech Pathologist	Zoomcast	Nampa	ID
13	955467	JENNIFER	SULLIVAN	jsullivan4r@lycos.co	Automation Specialis	Bluezoom	Bridgeport	CT
14	A3873	RONALD	GONZALES	rgonzales5p@apple.co	Automation Specialis	Shuffletag	Racine	WI
15	380630	VICTOR	COX	vcoxc9@virginia.edu	Librarian	Skalith	Bend	OR

Because you created this preparation using the **Add Preparation** button, you do not need to save anything. Every new preparation step is automatically saved.

Multiple preparations can be saved and created for a single dataset.

Remember that the original data from your dataset remains unchanged.

Manipulate data

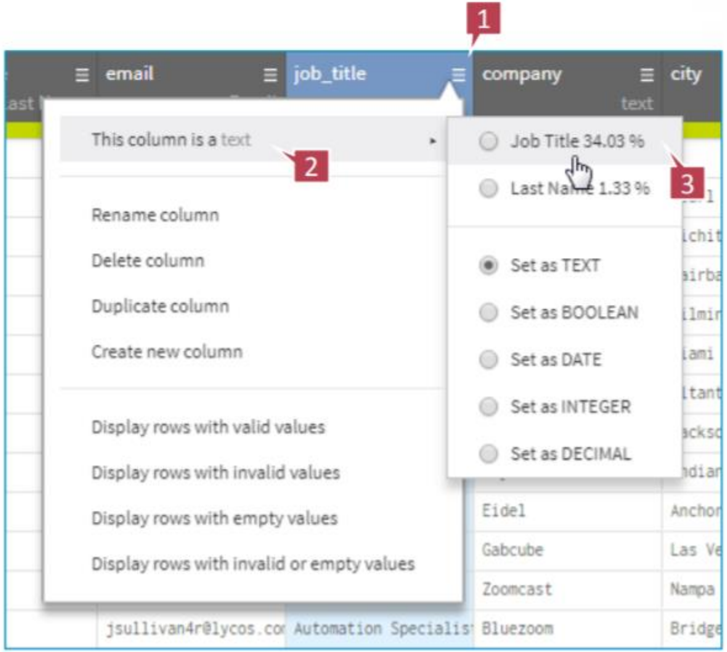
Semantic type

Talend Data Preparation automatically suggests the proper semantic type for each column of your datasets. It will help you to further discover the data. But you can change those suggestion at any time, based on your own experience.

The suggested semantic type for the **job_title** column is **Text**. Let’s change it to a more meaningful one, **Job Title** in this case.

1. Click the **menu icon** on the column header and select a new semantic type.
2. Point your mouse over **This column is a text**.
3. Select **Job Title** as new semantic type.

The Enterprise Edition of Talend Data Preparation allows you to create custom semantic types, as well as editing or removing the default ones.

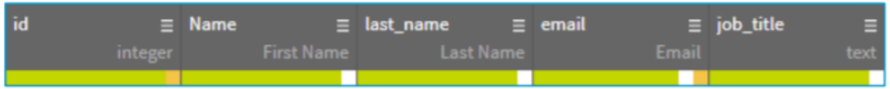


Manipulate data

Data quality bar

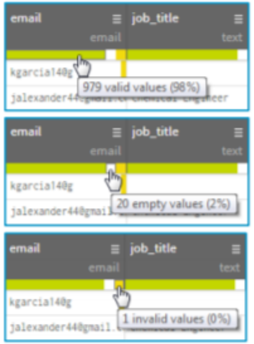
Under each column is a data quality bar that displays the amount of fields that have correct data, empty fields, or incorrect data. Each of these 3 are represented by a color.

- 🟩 **Green** – Data matches the cell format
- □ **White** – Empty cells
- 🟧 **Orange** – Data in the cell does not match the cell format



Let’s take a closer look at the quality bar for the **email** column. Exact numbers and percentages can be found by pointing your mouse over each color.

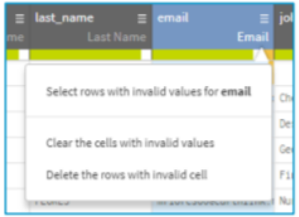
- 🟩 **Green** – 979 cells have data in the correct format
- □ **White** – 20 empty cells
- 🟧 **Orange** – 1 cells have entries in an incorrect format



979 valid values (98%)

20 empty values (2%)

1 invalid values (0%)



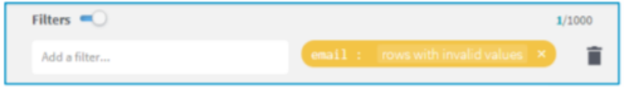
Select rows with invalid values for **email**

Clear the cells with invalid values

Delete the rows with invalid cell

Click any color to select, delete, or clear the cells with data in an invalid format. Click the orange section and click **Select rows with invalid values** for the **email** column to display the entries with an incorrect format.

Don’t forget to clear the filter to return to the full list.



email : rows with invalid values ×

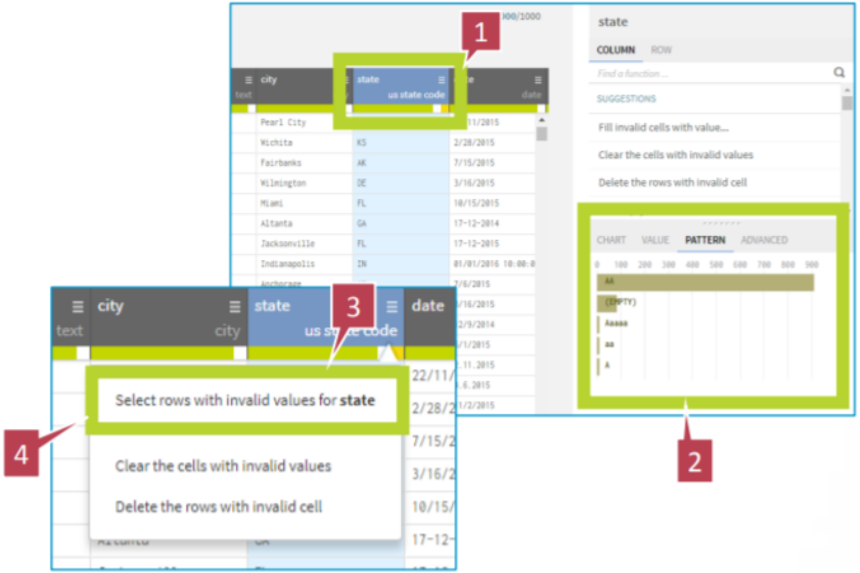
Manipulate data

Basic text manipulation

How to filter invalid rows:

1. Click the header of the **state** column.
2. In the bottom right is a **Pattern** table. Point your mouse over the rows to see counts. The top row indicates that 911 records contain a 2 letter state code. **You can click a bar to isolate those records (to remove the filter, click the x in the filter, or click the bin icon in the filter bar).**
3. On the data quality bar, click the **orange section**.
4. Click **Select rows with invalid values for state**.
5. 7 rows that contain invalid information will be displayed.

Here we are cleaning and changing the values in a field with invalid values. You will see how you can use the charts to help filter the data as well as change values directly in the grid.



city	state	date
Pearl City		11/2015
Wichita	KS	2/28/2015
Fairbanks	AK	7/15/2015
Wilmington	DE	3/16/2015
Miami	FL	10/15/2015
Altanta	GA	17-12-2014
Jacksonville	FL	17-12-2015
Indianapolis	IN	01/01/2016 10:00:0
Anchorage		7/6/2015

1
000/1000
city
text
state
us state code
date
state
COLUMN
ROW
Find a function...
SUGGESTIONS
Fill invalid cells with value...
Clear the cells with invalid values
Delete the rows with invalid cell
CHART
VALUE
PATTERN
ADVANCED
0
100
200
300
400
500
600
700
800
900
AA
(EMPTY)
Aaaaa
aa
A
2
3
city
state
us state code
date
4
Select rows with invalid values for state
Clear the cells with invalid values
Delete the rows with invalid cell

Manipulate data

Basic text manipulation

How to filter invalid rows:

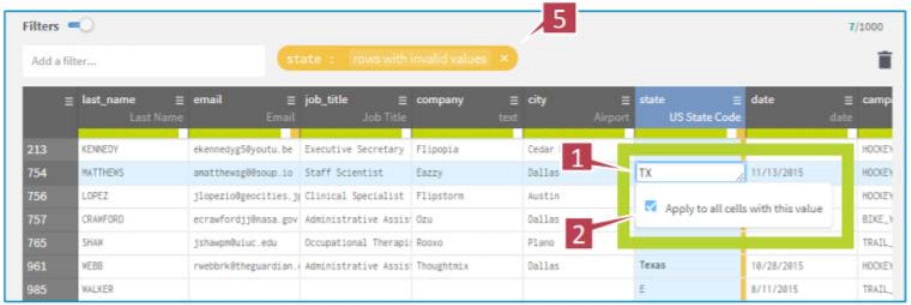
1. To edit the text value in a field, **double-click one of the cells** that contain **Texas**. Change **Texas** to **TX**. DO NOT **hit Enter yet!**

2. Under the cell that you are editing is a check box with the label **Apply to all cell with this value**. Check that box. NOW **hit Enter!** You have changed all cells with the value **Texas** to **TX**.

3. That should leave you with 2 rows with incorrect data. Check out the different functions and **you pick the one you want** to use to fix the invalid state codes!

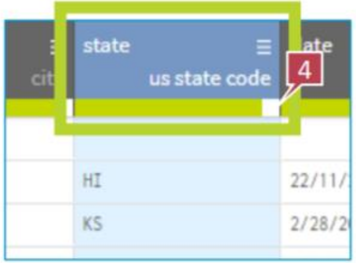
4. Once all actions and functions are applied, your **data quality bar** under the **state** column should now only contain **green and white**.

5. Click the **x** in the **state: rows with invalid values** filter to return to the full list.



	last_name Last Name	email Email	job_title Job Title	company text	city Airport	state US State Code	date date	camp
213	KENNEDY	ekennedyg5@youtu.be	Executive Secretary	Flipopia	Cedar			HOCKEY
754	MATTHEWS	amatthewsg0@soup.io	Staff Scientist	Eazzy	Dallas	TX	11/13/2015	HOCKEY
756	LOPEZ	jlopezio@geocities.j	Clinical Specialist	Flipstorm	Austin			HOCKEY
757	CRAWFORD	ecrawfordjj@nasa.gov	Administrative Assis	Ozu	Dallas			BIKE_
765	SHAW	jshawpm@uiuc.edu	Occupational Therapi	Rooxo	Plano			TRAIL_
961	WEBB	rwebbrk@theguardian.	Administrative Assis	Thoughtmix	Dallas	Texas	10/28/2015	HOCKEY
985	WALKER					E	8/11/2015	TRAIL_

☑ Apply to all cells with this value

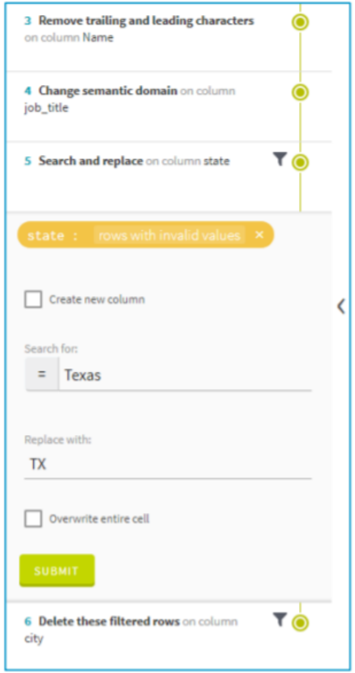


cit	state us state code	ate
	HI	22/11/
	KS	2/28/2

Manipulate data

Recipes

Each function that has been applied has been added to our recipe. Looking at the last steps in the recipe, it is easy to identify that we changed all fields that had **Texas** listed as a state to **TX**.



3 **Remove trailing and leading characters** on column Name

4 **Change semantic domain** on column job_title

5 **Search and replace** on column state

state : rows with invalid values ×

□ Create new column

Search for: = Texas

Replace with: TX

□ Overwrite entire cell

SUBMIT

6 **Delete these filtered rows** on column city

Manipulate data

Basic numeric manipulation

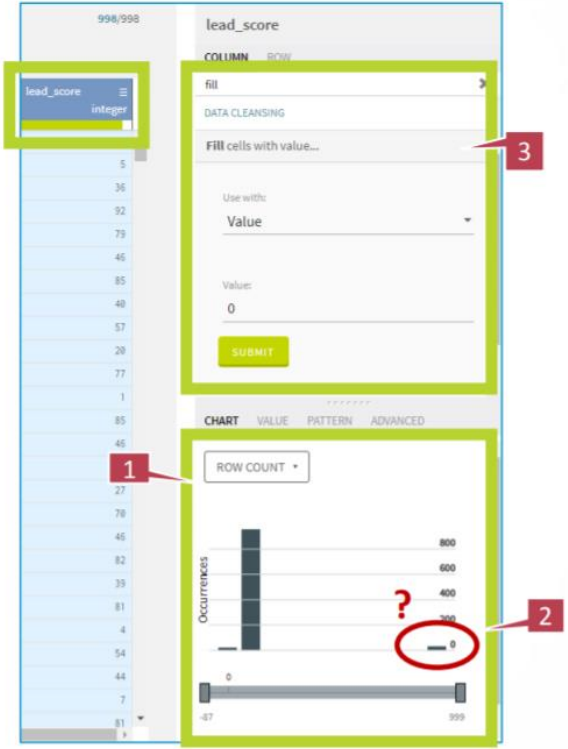
Next, let's look at the **lead_score** column.

1. Select the **lead_score** column. You will see that it contains basic integer values. But look at the **histogram graph** at the bottom right. The data is being skewed by some large value.

2. **Click the blue bar** on the far right of the graph. It should return 31 records with the value of 999. It looks like the default is set to 999. This time, to change the data, you will use a function called **Fill cell with value**.

3. Type **Fill** into the search field in the upper right. Click the function **Fill cell with value**. Set the value to **0** and click **Submit**.

Here, we are cleaning and changing outliers in a numeric field. You will see how you can use charts to help filter the data as well as change values directly in the data grid.



Use with: Value

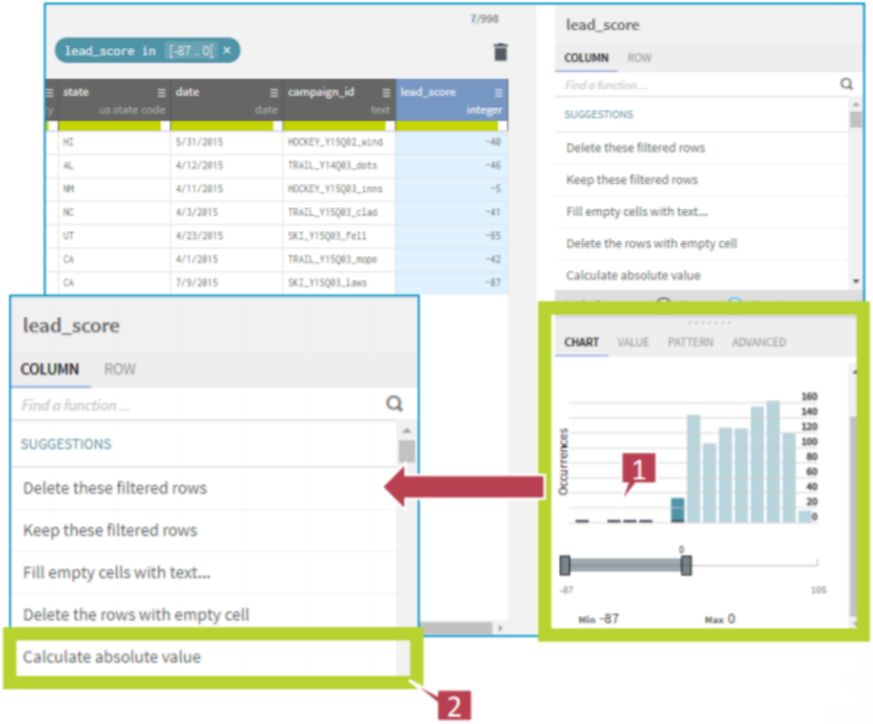
Value: 0

Manipulate data

Basic numeric manipulation

lead_score column, continued.

1. If you take a close look at the graph for the **lead_score** column, you will notice that there are negative lead scores.
2. Since we cannot have negative lead scores, let's remove those values. Under the suggested functions, click **Calculate absolute value**. This will keep the rows with their respective lead score numbers while dropping the negative sign.



lead_score in [-87 .. 0[

7/998

state us state code	date date	campaign_id text	lead_score integer
HI	5/31/2015	HOCKEY_Y15Q02_wind	-40
AL	4/12/2015	TRAIL_Y14Q03_dots	-46
NM	4/11/2015	HOCKEY_Y15Q03_inns	-5
NC	4/3/2015	TRAIL_Y15Q03_clad	-41
UT	4/23/2015	SKI_Y15Q03_fell	-65
CA	4/1/2015	TRAIL_Y15Q03_mope	-42
CA	7/9/2015	SKI_Y15Q03_laws	-87

Manipulate data

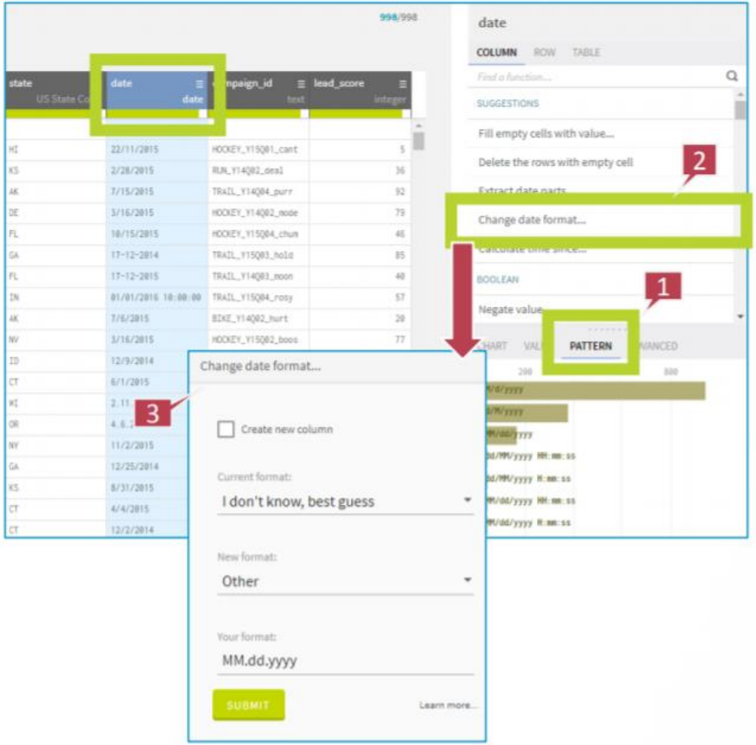
Date cleansing and formatting

Next, look at the **date** column.

1. Click the header of the **date** column, then change the view on the right to **Pattern**. This gives you a better view of the different date formats and masking used. Some dates are formatted in the European standards and others are formatted in the US standard. Some contain **-** and other **/**.

2. To standardize the dates, click **Change date format** under the suggested functions. Select a pre-existing format or type one in. Click **Submit** when ready.

How many times do we see a spreadsheet with all kinds of crazy data formats and standards? We all know that Excel can reformat a date field, but when the dates are a mix of European standards and US standard with different masking, Excel starts to break DOWN!



state US State Co	date date	mpaign_id text	lead_score integer
HI	22/11/2015	HOCKEY_Y15Q01_cant	5
KS	2/28/2015	RUN_Y14Q02_deal	36
AK	7/15/2015	TRAIL_Y14Q04_purr	92
DE	3/16/2015	HOCKEY_Y14Q02_mode	79
FL	10/15/2015	HOCKEY_Y15Q04_chum	46
GA	17-12-2014	TRAIL_Y15Q03_hold	85
FL	17-12-2015	TRAIL_Y14Q03_moon	40
IN	01/01/2016 10:00:00	TRAIL_Y15Q04_rosy	57
AK	7/6/2015	BIKE_Y14Q02_hurt	20
NV	3/16/2015	HOCKEY_Y15Q02_boos	77
ID	12/9/2014		
CT	6/1/2015		
WI	2.11		
OR	4.6.		
NY	11/2/2015		
GA	12/25/2014		
KS	8/31/2015		
CT	4/4/2015		
CT	12/2/2014		

Change date format...

☐ Create new column

Current format:
I don't know, best guess

New format:
Other

Your format:
MM.dd.yyyy

SUBMIT

Learn more...

Manipulate data

Date cleansing and formatting

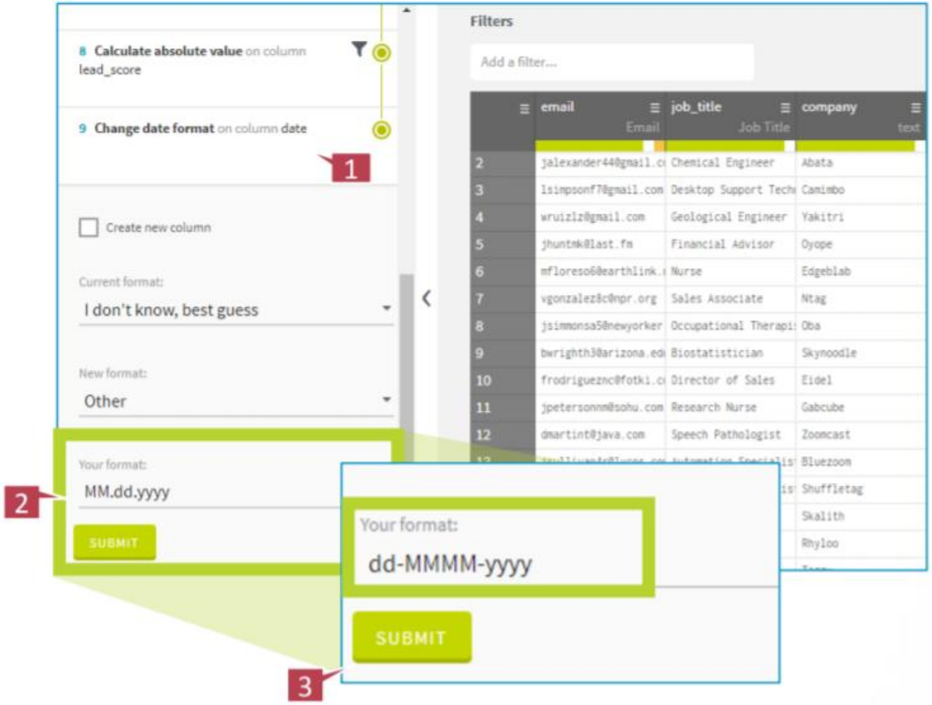
Modifying recipes is simple.

1. From the **recipe** on the left, highlight the last action.

2. In the drop-down for the **Change date format** operation, select **Other** (design your custom pattern). Enter **dd-MMMM-yyyy** (Date formatting is case sensitive so pay attention to the case).

3. Once you click **Submit**, the change will take effect. You can delete a step from the recipe list of actions on the left or click the green dot to inactivate that action.

4. You can also **reorder the steps of your recipe through drag & drop**. You will save time if you realize that a column you applied a function on, still does not fully contain the expected data.



8 Calculate absolute value on column lead_score

9 Change date format on column date

□ Create new column

Current format: I don't know, best guess

New format: Other

Your format: MM.dd.yyyy

Your format: dd-MMMM-yyyy

Filters

Add a filter...

	email Email	job_title Job Title	company text
2	jalexander44@gmail.c	Chemical Engineer	Abata
3	lsimpsonf7@gmail.com	Desktop Support Tech	Camimbo
4	wruiz1z@gmail.com	Geological Engineer	Yakitri
5	jhuntmk@last.fm	Financial Advisor	Oyope
6	mfloreso6@earthlink.	Nurse	Edgeblab
7	vgonzalez8c@npr.org	Sales Associate	Ntag
8	jsimmonsa5@newyorker	Occupational Therapi:	Oba
9	bwrighth3@arizona.ed	Biostatistician	Skynoodle
10	frodrigueznc@fotki.c	Director of Sales	Eidel
11	jpetersonnm@sohu.com	Research Nurse	Gabcube
12	dmartint@java.com	Speech Pathologist	Zoomcast
13	[illegible]	Automation Specialis	Bluezoom
			Shuffletag
			Skalith
			Rhyloo

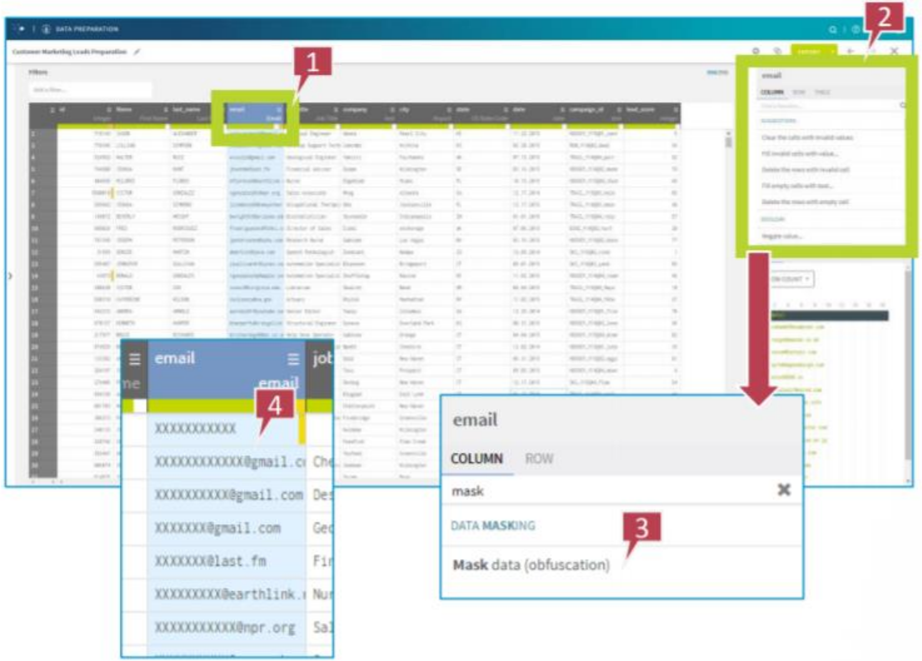
Manipulate data

Data masking

You can easily mask sensitive data.

1. Click the **email** column to select its content.
2. In the function list, search for **Mask data (Obfuscation)**.
3. Click it to apply the function on the email entries.
4. All the characters before @ are replaced by XXX, while the rest is left unchanged. This is the effect of the data masking function on entries whose semantic type is email. But the effects of the data masking will be different depending on a column's semantic type.

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, you will use the data masking function to generate functional substitutes.



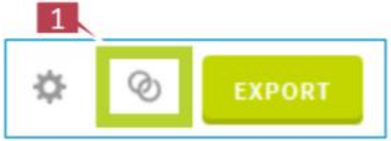
```
XXXXXXXXXXX
XXXXXXXXXXXX@gmail.c
XXXXXXXXXX@gmail.com
XXXXXXX@gmail.com
XXXXXXX@last.fm
XXXXXXXXX@earthlink.
XXXXXXXXXXX@npr.org
```

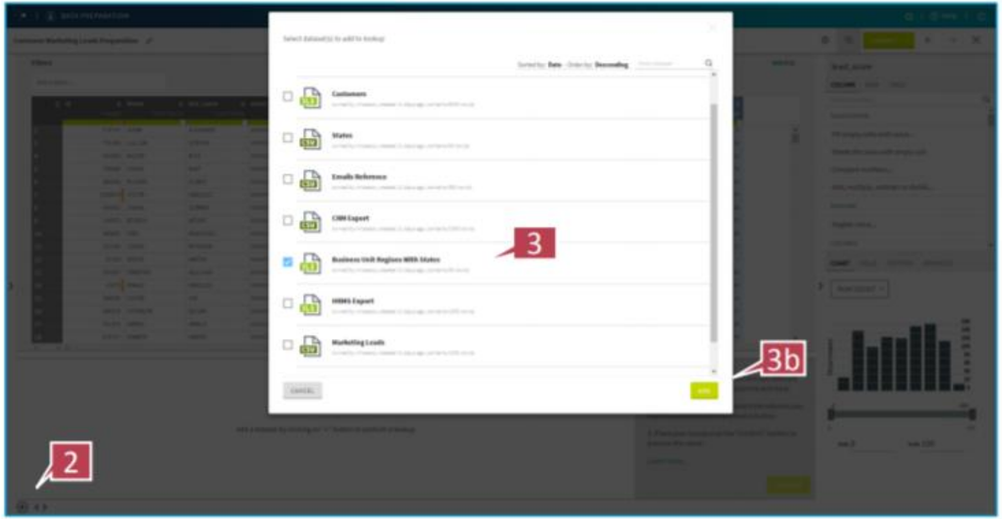
Manipulate data

Data blending

Data blending is about connecting data from different sources. It allows you to take data from another preloaded dataset and add them into the dataset you are currently working on.

1. Click the **Lookup** icon.
2. All the datasets that you have loaded plus some preloaded are available to choose from by clicking the **+** icon.
3. Click the checkbox in front of **Business Unit Regions With States** and click **Add**.




Select dataset(s) to add to lookup

Sorted by: Date · Order by: Descending

- ☐ Customers
- ☐ States
- ☐ Emails Reference
- ☐ CRM Export
- ☑ Business Unit Regions With States
- ☐ HRMS Export
- ☐ Marketing Leads

Manipulate data

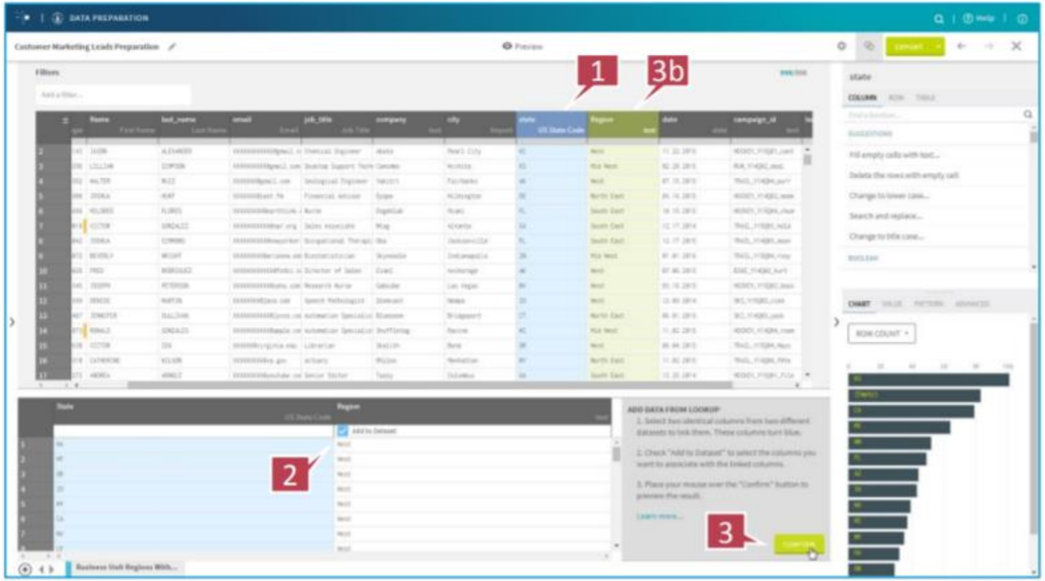
Data blending

Data blending, continued.

1. Click the **column you would like to blend**, the **state** column in your current dataset.

2. At the bottom, add the region information by clicking **Add to dataset** under the **Region** column header.

3. **Point your mouse over the Confirm button** to preview the changes, that are displayed in green. To accept the changes, click **Confirm**.



ADD DATA FROM LOOKUP

1. Select two identical columns from two different datasets to link them. These columns turn blue.

2. Check "Add to Dataset" to select the columns you want to associate with the linked columns.

3. Place your mouse over the "Confirm" button to preview the result.

Learn more...

Manipulate data

Group and standardize

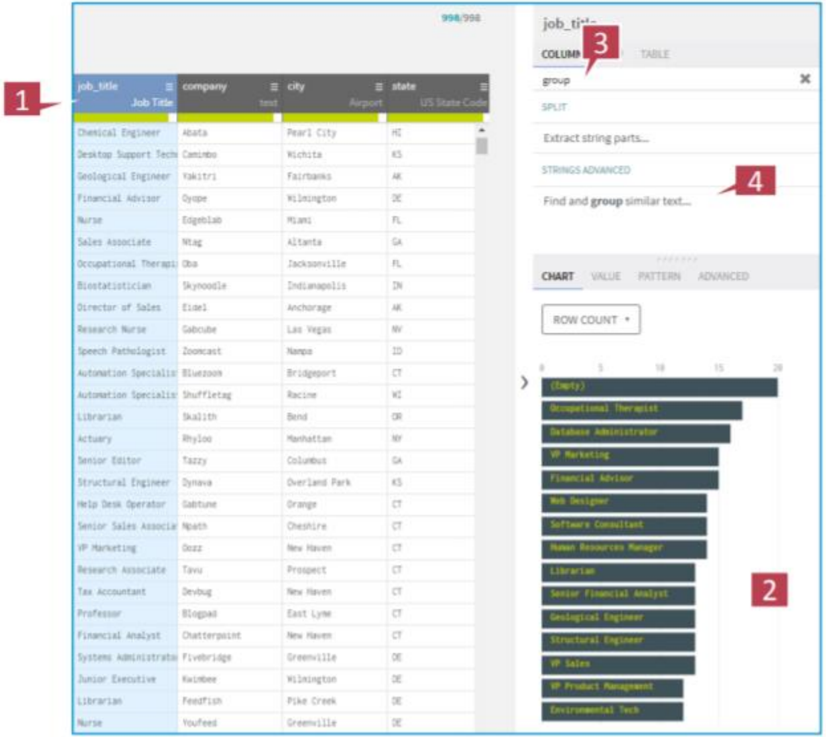
Group and standardize allows you to find cells that have similar content and group them together by changing the text to match.

1. Click the **job_title** column header.

2. The chart on the bottom right displays the large amount of slightly different job titles. To reduce the number of job titles, let's group similar job titles together.

3. In the search field, search for **group**.

4. Click the **Find and group similar text** function.



group

job_title	company	city	state
Job Title	text	Airport	US State Code
Chemical Engineer	Abata	Pearl City	HI
Desktop Support Tech	Camimbo	Wichita	KS
Geological Engineer	Yakitri	Fairbanks	AK
Financial Advisor	Oyope	Wilmington	DE
Nurse	Edgeblab	Miami	FL
Sales Associate	Ntag	Altanta	GA
Occupational Therapi	Oba	Jacksonville	FL
Biostatistician	Skynoodle	Indianapolis	IN
Director of Sales	Eidel	Anchorage	AK
Research Nurse	Gabcube	Las Vegas	NV
Speech Pathologist	Zoomcast	Nampa	ID
Automation Specialis	Bluezoom	Bridgeport	CT
Automation Specialis	Shuffletag	Racine	WI
Librarian	Skalith	Bend	OR
Actuary	Rhyloo	Manhattan	NY
Senior Editor	Tazzy	Columbus	GA
Structural Engineer	Dynava	Overland Park	KS
Help Desk Operator	Gabtune	Orange	CT
Senior Sales Associa	Npath	Cheshire	CT
VP Marketing	Oozz	New Haven	CT
Research Associate	Tavu	Prospect	CT
Tax Accountant	Devbug	New Haven	CT
Professor	Blogpad	East Lyme	CT
Financial Analyst	Chatterpoint	New Haven	CT
Systems Administrato	Fivebridge	Greenville	DE
Junior Executive	Kwimbee	Wilmington	DE
Librarian	Feedfish	Pike Creek	DE
Nurse	Youfeed	Greenville	DE

Manipulate data

Find and group similar text

Group and standardize, continued.

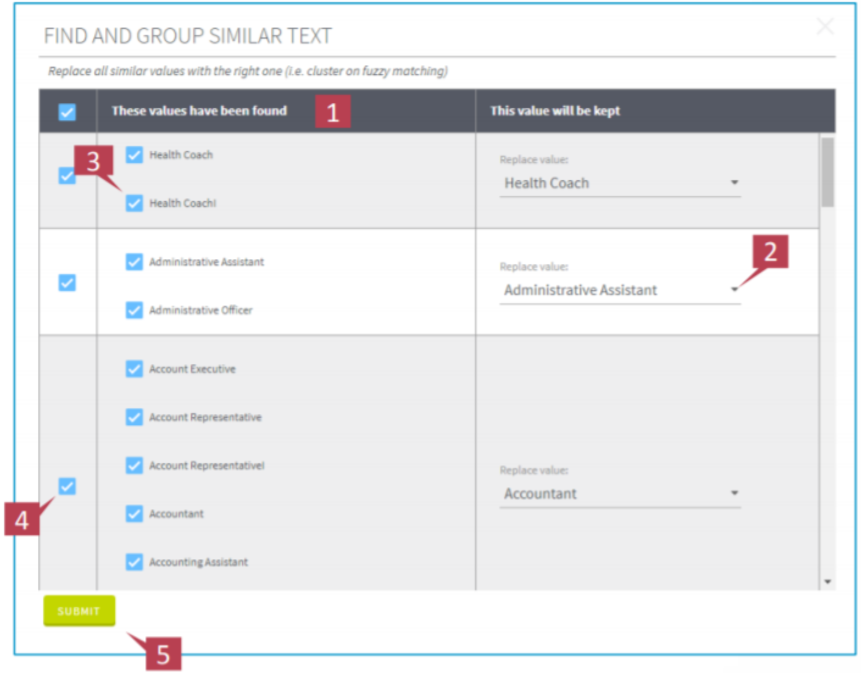
1. All similar job titles are grouped together in the second column

2. The third column suggests a job title that could **replace** the data in the second column. You can **use the drop-down list to choose a different job title or type in an appropriate job title**.

3. If you do not want to change a specific job title, leave the check box in front of the job title **clear**.

4. If you do no want to change a group of job titles, leave the check box in front of the first column **clear**.

5. Click **Submit** when finished.



FIND AND GROUP SIMILAR TEXT

Replace all similar values with the right one (i.e. cluster on fuzzy matching)

☑	These values have been found	This value will be kept
☑	☑ Health Coach ☑ Health Coachl	Replace value: Health Coach ▾
☑	☑ Administrative Assistant ☑ Administrative Officer	Replace value: Administrative Assistant ▾
☑	☑ Account Executive ☑ Account Representative ☑ Account Representativel ☑ Accountant ☑ Accounting Assistant	Replace value: Accountant ▾