



Talend Data Preparation Free Desktop

Getting Started Guide
V2.1





Talend Data Preparation Training

Getting Started Guide

To navigate to a specific location within this guide, click one of the boxes below.

**Overview of
Data
Preparation**

**Access Data
Preparation
And Getting
Started**

**Simple
Cleansing**

**Basic Data
Manipulation**

**Date
Cleansing and
Formatting**

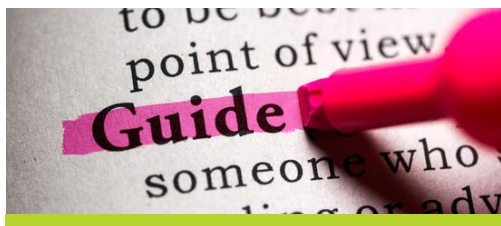
Talend Data Preparation Training

Getting Started Guide

[Overview of Data Preparation](#)[Access Talend Data Preparation And Getting Started](#)[Simple Cleansing Examples](#)[Basic Data Manipulation](#)[Date Cleansing and Formatting](#)

About this guide

What is this Talend Training Guide?



Using the Talend Data Preparation

Guide provides step-by-step instructions to build and run an end-to-end Data Preparation Recipe.



The demo is built on a real world use-case Marketing Data. The data is reflective of what many users deal with every day but in spreadsheets and they are forced to use complex macros or, worse, VBA Scripts.



The goal is to have you using the easy-to-use web UI tool. Understand how Talend can be used to discover, cleanse, format and enrich your data.

About Talend Data Preparation

The self-service data prep tool for everyone



Watch the [Introduction to Talend Data Preparation Video](#)



Create clean and valuable data in minutes, not hours

- Single point of access across data sources
- Interactive discovery, cleansing and formatting
- Automate and reuse data preparation tasks



Self-service data cleansing for everyone

- Put data at work for your daily tasks
- Auto-discover and browse your data
- Get guided on your way to actionable data



Empower business and IT towards faster business insights

- Managed self service data access across the enterprise
- Prevent data inconsistencies and leaks that adversely affect the business
- Free the IT backlog from mundane tasks and improve productivity

If you have already installed Talend Data Preparation, [click here](#) to skip over the installation instructions.

Marketing Lead Data Preparation

Data in the file “customer marketing leads.csv” is marketing lead contact information from a business partner and has data quality issues or data that needs to be reformatted in several fields. Analyzing the raw data in this file would lead to poor results due to incorrect data or missing values. Fixing this using Excel would take hours.

In this demo, we will guide you through some beginner Data Preparation actions which can be challenging in Excel.

You will experience:

- Quickly changing data values after identifying them through cool graphs and filters with no coding!
- Check out great features like histogram-graphs and how they can help fix your data!
- How to manipulate text, dates, and numeric data in a file with just a few clicks!

customer marketing leads - Microsoft Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	id	Name	last_name	email	job_title	company	city	state	date	campaign	lead_score												
2	771396	Kathryn	Garcia	kgarcia14@g																			
3	718143	Jason	Alexander	jalexander	Chemical f	Abata	Pearl City	HI	#####	HOCKEY_Y	5												
4	770396	Lillian	Simpson	lsimpson7	Desktop St	Camimbo	Wichita	KS	#####	RUN_Y14C	36												
5	524952	WALTER	Ruiz	wruizl@g	Geological	Yakitri	Fairbanks	AK	#####	TRAIL_Y14	92												
6	744980	Joshua	Hunt	jhuntmk@	Financial A	Oyope	Wilmington	DE	#####	HOCKEY_Y	79												
7	404656	Mildred	Flores	mflores06	Nurse	Edgeblab	Miami	FL	#####	HOCKEY_Y	46												
8	958@018	Victor	Gonzalez	vgonzalez	Sales Asso	Ntag	Altanta	GA	#####	TRAIL_Y15	85												
9	595042	Joshua	Simmons	jsimmons	Occupatio	Oba	Jacksonvill	FL	#####	TRAIL_Y14	40												
10	149072	Beverly	Wright	bwright3i	Biostatistic	Skynoodle	Indianapol	IN	8/6/2015	TRAIL_Y15	57												
11	609026	Fred	Rodriguez	frodriquez	Director o	Eidel	Anchorage	AK	7/6/2015	BIKE_Y14C	20												
12	761545	Joseph	Peterson	jpetersonn	Research f	Gabcube	Las Vegas	NV	#####	HOCKEY_Y	77												
13	31599	Denise	Martin	dmartin@	Speech Pa	Zoomcast	Nampa	ID	#####	SKI_Y150C	1												
14	955467	Jennifer	Sullivan	jsullivan4r	Automatic	Bluezoom	Bridgeport	CT	6/1/2015	SKI_Y140C	85												
15	A3873	Ronald	Gonzales	rgonzales5	Automatic	Shuffletag	Racine	WI	#####	HOCKEY_Y	46												
16	380630	VICTOR	Cox	vcocx9@v	Librarian	Skalith	Bend	OR	4/6/2015	TRAIL_Y15	10												
17	690310	Catherine	Wilson	cwilsonca	Actuary	Rhyloo	Manhattar	NY	#####	TRAIL_Y14	27												
18	542272	Andrea	Arnold	aarnoldfr	Senior Edit	Tazzy	Columbus	GA	#####	HOCKEY_Y	70												
19	678157	Kenneth	Harper	kharperrf	Structural	Dynava	Overland	KS	#####	HOCKEY_Y	46												
20	217977	Bruce	Richards	brichardsg	Help Desk	Gabtune	Orange	CT	4/4/2015	HOCKEY_Y	82												
21	874929	Brandon	Porter	bportergr	Senior Sale	Npath	Cheshire	CT	#####	HOCKEY_Y	39												
22	133382	Angela	Stone	astonehb	VP Market	Oozz	New Have	CT	#####	HOCKEY_Y	81												
23	254197	Judith	Bell	jbellhq@w	Research /	Tavu	Prospect	CT	9/3/2015	HOCKEY_Y	4												
24	279406	Peter	Torres	ptorreskk	Tax Accou	Devbug	New Have	CT	6/3/2015	SKI_Y150C	54												
25	894168	Alan	Ryan	aryanlh@	Professor	Blogpad	East Lyme	CT	#####	TRAIL_Y14	44												
26	801703	Mark	Garcia	mgarciahr	Financial A	Chatterpol	New Have	CT	#####	HOCKEY_Y	7												
27	306273	Paul	Bishop	pbishop2n	Systems A	Fivebridge	Greenville	DE	#####	HOCKEY_Y	81												
28	240133	Juan	Ford	jford3p@	c Junior Exe	Kwimbee	Wilmington	DE	#####	HOCKEY_Y	30												
29	228742	Christophe	Larson	clarson5a	Librarian	Feedfish	Pike Creek	DE	8/4/2015	TRAIL_Y15	6												
30	362447	Andrea	Lewis	alewis5c	Nurse	Youfeed	Greenville	DE	4/9/2015	HOCKEY_Y	39												
31	806874	Jennifer	Gibson	jgibsoncb	Human Re	Jaxbean	Wilmington	DE	#####	HOCKEY_Y	80												
32	814025	Joshua	Hernandez	jhernandez	Nurse	PracTwim	Bear	DE	#####	TRAIL_Y14	16												

customer marketing leads

Requirements For Talend Data Preparation

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

Hardware requirements:

Processor	64-bit processor is required (Note: 32-bit is not supported)
Allocated memory	1GB minimum
Disk space	500MB minimum free disk space

Software requirements:

Operating system	<ul style="list-style-type: none">Windows 7 or more recentMac OS X 10.7 “Lion” or more recent
------------------	--

Compatible web browsers:

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

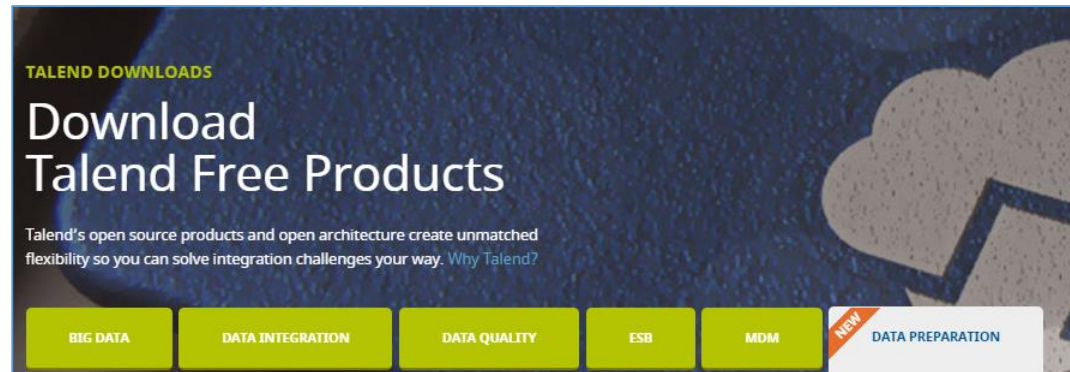
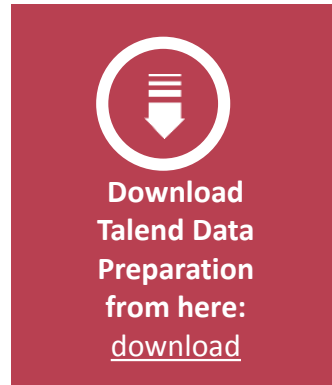
Java:

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.

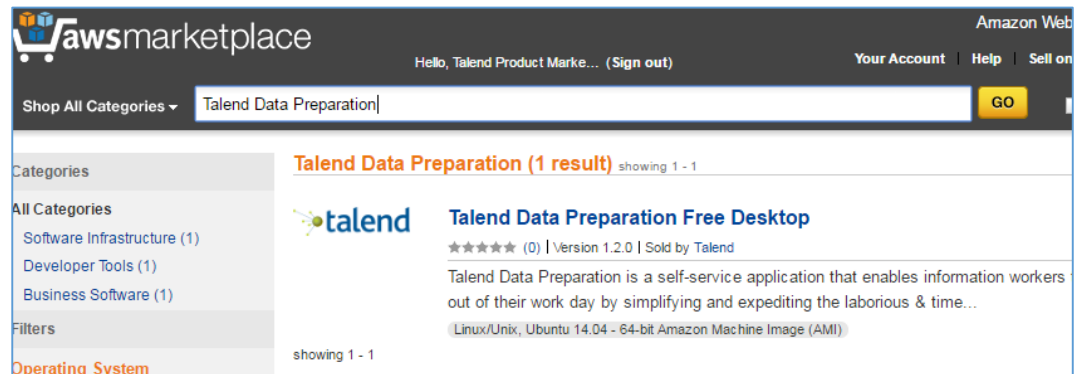
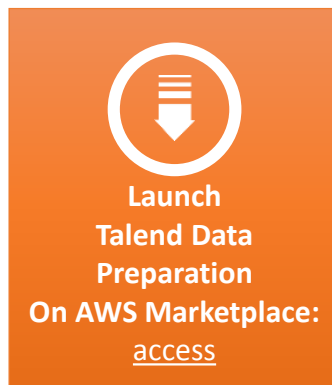
How do I get started with Talend Data Preparation?

You can use Talend Data Preparation:

- Either on a **On-Premise Mode** by downloading Talend Data Preparation on your Desktop



- Or on a **On-Demand Mode** by launching Talend Data Preparation on an Amazon Machine Image (AMI) through AWS Marketplace Cloud Provider



How do I download Talend Data Preparation?



Download
Talend Data
Preparation
[here](#)

TALEND DOWNLOADS

Download Talend Free Products

Talend's open source products and open architecture create unmatched flexibility so you can solve integration challenges your way. [Why Talend?](#)

BIG DATA

DATA INTEGRATION

DATA QUALITY

ESB

MDM

NEW DATA PREPARATION

Data Preparation Free Desktop

Tired of tearing your hair out to get clean, usable data? Download Talend Data Preparation, a free desktop application that does the work for you.

Talend Data Preparation Free Desktop
Version 1.2.0 | Basic Data Preparation

WINDOWS

MAC

Features

- Single user, desktop-based application
- Fully functional data preparation capabilities
- Import, export, and merge Excel and CSV files
- Auto-discovery, profiling, smart suggestions, and data visualization
- Cleansing and enrichment functions
- Completely free

Need an enterprise solution?

Try **Talend Data Preparation subscription version** to manage data access and facilitate collaboration across the enterprise.

Subscription-based licensing per user

Additional Features

- Multi-user, role-based access
- Collaborate and share data prep recipes
- Operationalize into any data integration flow
- Support for hundreds of data sources and targets
- Web and email support

REQUEST INFO

[Learn why you should upgrade to our commercial editions](#)

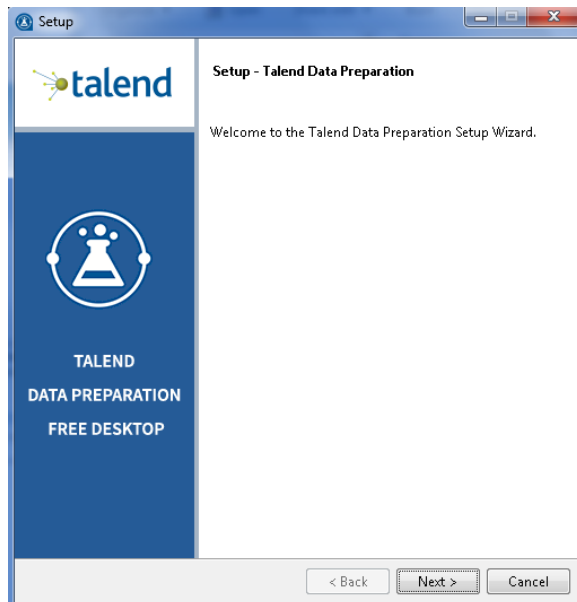
How do I set-up Talend Data Preparation if I am a Windows user?



The Windows version is provided as a standard Microsoft Windows Installer. It will require local administrator rights. Follow the steps below to install and start Data Preparation:

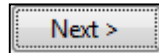
1

Locate the downloaded file from above and double click **Talend-DataPreparation-Free-Desktop-2.1.exe**



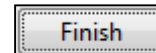
2

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.



Alternate installation for Windows users



If you do not have the local administrator rights required to use the installer, follow the steps below to install Data Preparation via a .zip file:

1

In the Talend Data preparation download page, scroll down to the **Other Releases** section.

2

Download the **Talend-DataPreparation-Free-Desktop-windows-2.1.0.zip** file.

3

Unzip the file on any location of your drive.

4

Run the **.exe** file to begin using the Talend Data Preparation tool.

Other Releases

Version	Release Date	File Name	Release Type	Supported Operating Systems	Size	Mirror
1.3.0	September 30, 2016	Talend-DataPreparation-Free-Desktop-1.3.0.exe	Main	Windows	179MB	US Europe
1.3.0	September 30, 2016	Talend-DataPreparation-Free-Desktop-1.3.0-apache.exe	Main	Windows	104MB	US Europe
1.3.0	September 30, 2016	Talend-DataPreparation-Free-Desktop-1.3.0-apache.dmg	Main	MAC	98MB	US Europe
1.3.0	September 30, 2016	Talend-DataPreparation-Free-Desktop-windows-1.3.0.zip	Main	Windows	241MB	US Europe

How do I set-up Talend Data Preparation if I am a Mac OS X user?



Follow the steps below to install and start Data Preparation:

1

Double-click the Talend-DataPreparation-Free-Desktop-2.1.dmg file to open the package.

2

Drag and drop into the Applications folder.

3

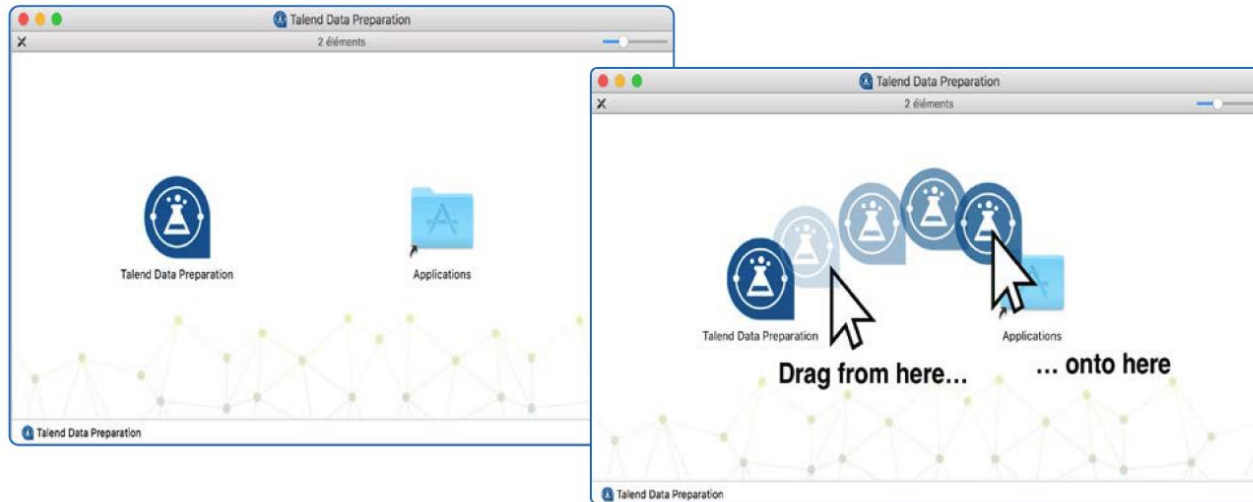
Talend Data Preparation will now be in you list of your Applications, locate the icon and open by double-clicking.

4

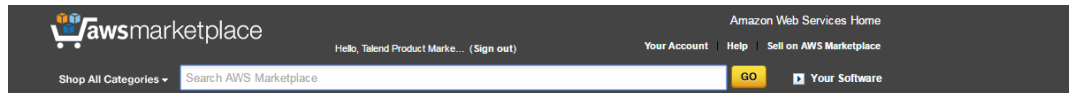
To disable App Nap and ensure optimal performance, follow this quick procedure:

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

2. Enter the following command:
`defaults write org.talend.dataprep NSAppSleepDisabled -bool YES`



How do I get started with Talend Data Preparation?



Talend Data Preparation Free Desktop

Sold by: Talend | See product video [📺](#)

Talend Data Preparation is a self-service application that enables information workers to cut hours out of their work day by simplifying and expediting the laborious & time consuming process of preparing data for analysis or other data-driven tasks. It enables anyone to quickly discover, cleanse, standardize, shape and enrich data using in-memory technologies, data visualization, and smart guidance. Talend Data Preparation reduces from hours to minutes the time it takes to get clean, useful data from Excel and CSV files into your favorite data analysis tool. It allows anyone to put... [Read more](#)

Customer Rating	★★★★★ (0 Customer Reviews)
Latest Version	1.2.0
Operating System	Linux/Unix, Ubuntu 14.04
Delivery Method	64-bit Amazon Machine Image (AMI) (Read more)
Support	See details below
AWS Services Required	AmazonEC2, AmazonEBS, AmazonS3, Cloud Formation
Highlights	<ul style="list-style-type: none"> Create clean and valuable data in minutes, not hours: click here to access the getting started content: https://www.talend.com/download/talend-open-studio#t8-gs Self-service data cleansing for anyone Empower business and IT towards faster business insights. Click here to learn more about the commercial version: http://www.talend.com/products/data-preparation.

Product Description

Talend Data Preparation is a self-service application that enables information workers to cut hours out of their work day by simplifying and expediting the laborious & time consuming process of preparing data for analysis or other data-driven tasks. It enables anyone to quickly discover, cleanse, standardize, shape and enrich data using in-memory technologies, data visualization, and smart guidance. Talend Data Preparation reduces from hours to minutes the time it takes to get clean, useful data from Excel and CSV files into your favorite data analysis tool. It allows anyone to put data at work and to reuse the data clean up recipes whenever data is updated and eliminates rework.

<div>Continue</div> <p>You will have an opportunity to review your order before launching or being charged.</p>			
Pricing Details			
For Region US East (N. Virginia)			
Hourly Fees Total hourly fees will vary by instance type and EC2 region.			
EC2 Instance Type	Software	EC2	Total
t2.medium	\$0.00/hr	\$0.052/hr	\$0.052/hr
m3.medium	\$0.00/hr	\$0.067/hr	\$0.067/hr
m3.large	\$0.00/hr	\$0.133/hr	\$0.133/hr
m3.xlarge	\$0.00/hr	\$0.266/hr	\$0.266/hr
c3.4xlarge	\$0.00/hr	\$2.10/hr	\$2.10/hr
cr1.8xlarge	\$0.00/hr	\$3.50/hr	\$3.50/hr
hi1.4xlarge	\$0.00/hr	\$3.10/hr	\$3.10/hr
hs1.8xlarge	\$0.00/hr	\$4.60/hr	\$4.60/hr
g2.2xlarge	\$0.00/hr	\$0.65/hr	\$0.65/hr
c3.8xlarge	\$0.00/hr	\$1.68/hr	\$1.68/hr
i2.xlarge	\$0.00/hr	\$0.853/hr	\$0.853/hr
i2.2xlarge	\$0.00/hr	\$1.705/hr	\$1.705/hr
i2.4xlarge	\$0.00/hr	\$3.41/hr	\$3.41/hr
i2.8xlarge	\$0.00/hr	\$6.82/hr	\$6.82/hr
r3.large	\$0.00/hr	\$0.166/hr	\$0.166/hr
r3.xlarge	\$0.00/hr	\$0.333/hr	\$0.333/hr
r3.2xlarge	\$0.00/hr	\$0.665/hr	\$0.665/hr
r3.4xlarge	\$0.00/hr	\$1.33/hr	\$1.33/hr
r3.8xlarge	\$0.00/hr	\$2.66/hr	\$2.66/hr
c4.large	\$0.00/hr	\$0.105/hr	\$0.105/hr



Launch
Talend Data
Preparation
On AWS Marketplace:
[access](#)

To **launch Talend Data Preparation on AWS Marketplace**, you need to:

- Log in with or create your AWS account
- Confirm your subscription: the software is free of charge, you will be charged for AWS Infrastructure Usage
- Configure and launch an AMI instance
- Connect to the instance
- Launch the software

Main page – Preparations and datasets

When the application is started, the first page to be displayed will be the “Preparations” view:

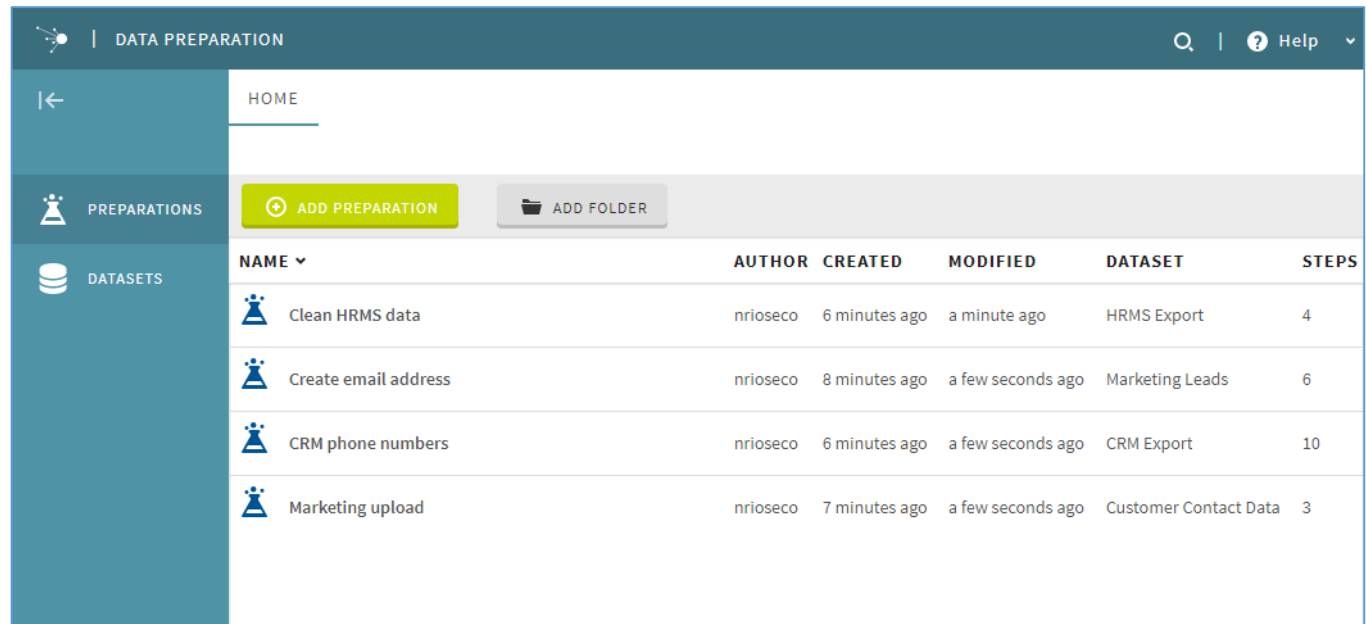
This view displays the list of all the preparations you have worked on. A preparation is the outcome of the different steps applied to cleanse your data. You can export this outcome as a file. A preparation takes one dataset and applies a recipe to produce the final result. The original data is never altered.





From this page, you can also access the “Datasets” view:

Displayed here will be a list of all the current datasets you have been working on or have imported. Datasets can be a local or remote file that can be imported into the Talend Data Preparation Tool (or from a database connection or other data sources, although not in the context of the Free Desktop version). A dataset is used as raw material for one or more preparations.

From this page, you can:

- Add new preparations
- Organize your preparations into folders
- Import and create new datasets
- Star favorite datasets

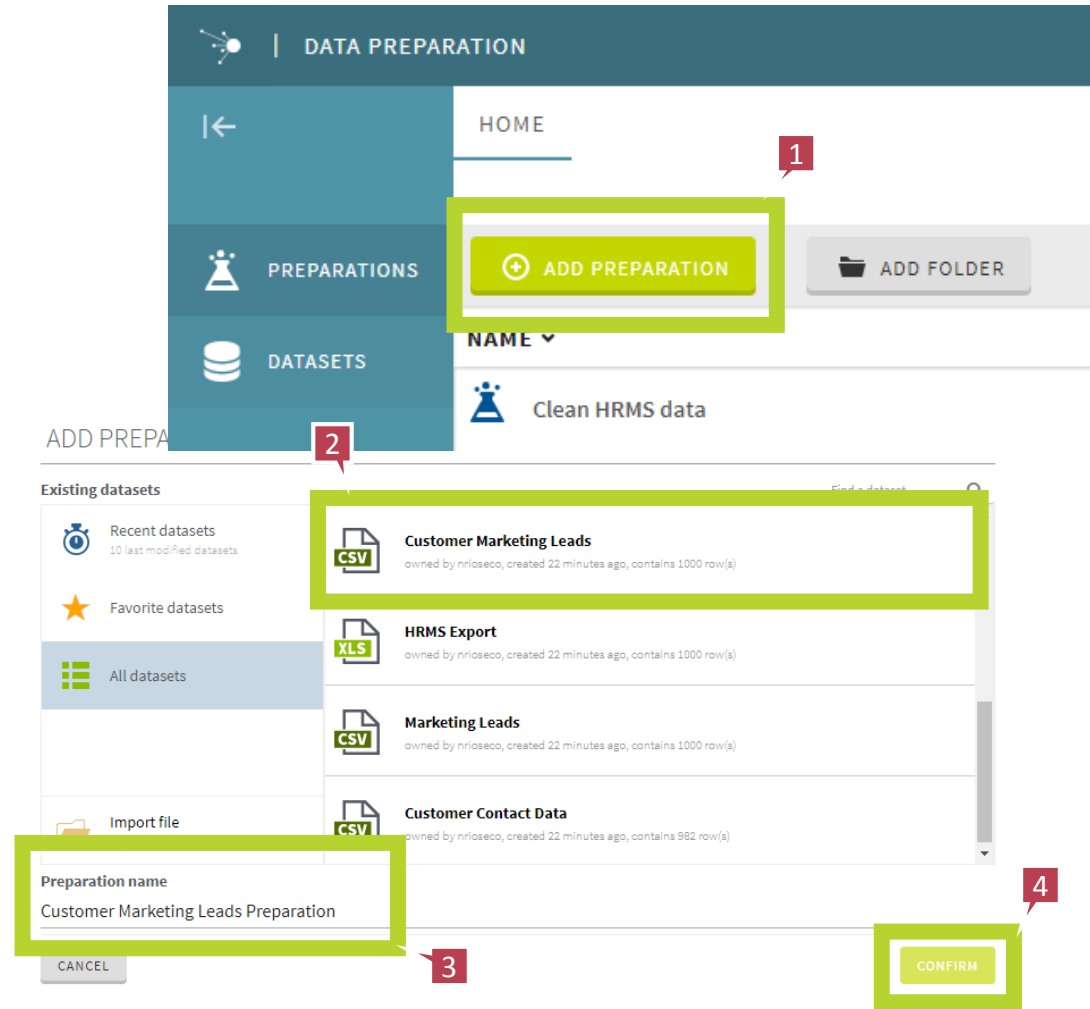


NAME		AUTHOR	CREATED	MODIFIED	DATASET	STEPS
	Clean HRMS data	nrioseco	6 minutes ago	a minute ago	HRMS Export	4
	Create email address	nrioseco	8 minutes ago	a few seconds ago	Marketing Leads	6
	CRM phone numbers	nrioseco	6 minutes ago	a few seconds ago	CRM Export	10
	Marketing upload	nrioseco	7 minutes ago	a few seconds ago	Customer Contact Data	3

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.



Data Preparation Guided Tour

In this demo,
we will demonstrate how to ...

**Do Simple
Cleansing Exercises**

Manipulate Data

**Standardize and
Enrich Data**

Simple Cleansing Examples

To execute basic formatting and Cleansing:

First, let's fix the Name Field:

1. Navigate to the column titled **NAME** and click the header.
2. 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.
3. You may need to Scroll down depending on the screen resolution to find the function **"Change Style to Title Case"**. Hover over the function to preview how the data will change. Click the function to apply the changes.

The screenshot shows the Talend Data Preparation interface for a project titled 'Customer Marketing Leads Preparation'. The main data table has columns: id, first name, last name, email, job_title, company, city, state, us state code, date, campaign_id, and lead_score. The 'Name' column is highlighted, and its header is clicked (indicated by a red arrow labeled '1').

In the top-right corner, a box labeled 'Name' contains a list of functions. A red arrow labeled '2' points to this box. The functions listed include: Fill empty cells with text..., Delete the rows with empty cell, Mask data (obfuscation), Change to upper case, Change to lower case, Negate value, Concatenate with..., and F COUNT.

A zoomed-in view of the 'Name' column shows the data being cleaned. The 'Name' column is highlighted, and its header is clicked (indicated by a red arrow labeled '1'). The data in the 'Name' column is shown in a table with columns 'id' and 'Name'. The data is as follows:

id	Name
1	771396 Kathryn
2	718143 Jason
3	770396 Lillian
4	524952 WALTER
5	744980 Joshua
6	404656 Mildred
7	958018 Victor
8	595042 Joshua

A red arrow labeled '3b' points to the 'Name' column header. A red arrow labeled '3' points to the 'Change to title case' function in the 'Name' box.

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 trailing Spaces and the last name has been defined as an incorrect type.

Simple Cleansing Examples

To execute the basic formatting and cleansing:

NAME column, continued

1. While looking at the data you will see white boxes in front of or behind some names, for example "Joshua".
2. To remove the white boxes, search for and select the function "Remove trailing and leading characters" and click "Submit".

The screenshot shows the Talend Data Preparation interface. The main window displays a table titled 'Customer Marketing Leads Preparation' with columns: id, Name, last_name, email, job_title, company, city, state, date, campaign_id, and lead_score. The 'Name' column contains various names, some with white boxes indicating trailing or leading characters. A red arrow points to the 'Joshua' entry in the 'Name' column. A green box highlights the 'Name' column header. A second green box highlights the 'Remove trailing and leading characters...' suggestion in the 'Name' column's context menu. A third green box highlights the 'SUGGESTIONS' section of the context menu.

id	Name	last_name	email	job_title	company	city	state	date	campaign_id	lead_score
1	771396 Kathryn	Garcia	kgarcia14@gmail.com	Chemical Engineer	Abata	Pearl City	HI	22/11/2015	HOOKEY_Y15001_cant	5
2	718143 Jason	Alexander	jalexander44@gmail.com	Desktop Support Tech	Cantembo	Wichita	KS	2/28/2015	RUK_Y14002_deal	36
3	778396 Lillian	Slapson	lslapson7@gmail.com	Geological Engineer	Yakitri	Fairbanks	AK	7/15/2015	TRAIL_Y14004_purr	92
4	524952 WALTER	Rutz	wrutzel@gmail.com	Financial Advisor	Oyope	Wilmington	DE	3/16/2015	HOOKEY_Y14002_node	79
5	744980 Joshua	Hunt	jhuntelast@earthlink.net	Nurse	Edgelab	Miami	FL	10/15/2015	HOOKEY_Y15004_shum	46
6	404656 Mildred	Flores	mflores08@earthlink.net	Sales Associate	Wag	Atlanta	GA	17-12-2014	TRAIL_Y15003_hold	85
7	958018 Victor	Gonzalez	vgonzalez8@chp.org	Occupational Therapist	Oha	Jacksonville	FL	17-12-2015	TRAIL_Y14003_moon	40
8	595042 Joshua	Simmons	jsimmons0@brevyorker.com	Biostatistician	Sky noodle	Indianapolis	IN	01/01/2016 10:00:00	TRAIL_Y15004_rossy	57
9	149072 Beverly	Wright	bwright3@arizona.edu	Director of Sales	Eidel	Anchorage	AK	7/6/2015	BKE_Y14002_hurt	20
10	609026 Fred	Rodriguez	frrodriguez@fotki.com	Research Nurse	Gabcube	Las Vegas	NV	3/16/2015	HOOKEY_Y15002_boos	77
11	761545 Joseph	Peterson	jpeterson@bnu.com	Speech Pathologist	Zoomcast	Nampa	ID	12/9/2014	SKL_Y15002_vied	1
12	31599 Denis	Martin	dmartin@java.com	Automation Specialist	Bluezoom	Bridgeport	CT	6/1/2015	SKL_Y14003_yack	85
13	955467 Dennis	Gonzales	dgonzales@people.com	Librarian	Shufflatag	Racine	WI	2-11-2015	HOOKEY_Y14004_roam	46
14	388630 VICTOR	Cox	vcoc@virginia.edu	Senior Editor	Tazzy	Columbus	GA	12/25/2014	HOOKEY_Y15001_fille	70
15	698318 Cathie	Wilson	cwilsonca@va.gov	Structural Engineer	Dynava	Overland Park	KS	8/31/2015	HOOKEY_Y14003_ione	46
16	542272 Andry	Arnold	aarnold@f@youtube.com	Help Desk Operator	Gabtune	Orange	CT	4/4/2015	HOOKEY_Y15004_nine	82
17	576157 Kenna	Harper	kharpert@calipolis.com	Senior Sales Associate	Nath	Cheshire	CT	12/2/2014	HOOKEY_Y15001_tute	39
18	217977 Bruce	Porter	bporter@offena.gov	VP Marketing	Oozz	New Haven	CT	5/31/2015	HOOKEY_Y15002_eggs	81
19	874029 Brando	Stone	astone@whitehouse.gov							
20	133382 Angus									
21										

Recipes

1. After each function is selected, it is added to the Recipe panel on the left.
2. To delete a recipe line item, hover over the line item and click the trash can.
3. To rename a preparation, click the pencil icon and enter the new name.
4. The recipe panel can be hidden by clicking the arrow.
5. To export the result of your preparation, click export then select the file type.

The screenshot shows the Talend Data Preparation interface. The recipe panel on the left is titled 'Customer Marketing Leads Preparation'. A red callout '1' points to the recipe panel. A red callout '2' points to the 'Remove step' button. A red callout '3' points to the 'Add step' button. A red callout '4' points to the arrow icon to hide the recipe panel. A red callout '5' points to the 'EXPORT' button. The data table below shows a list of customer marketing leads with columns: id, Name (first name, last name), email, job_title, company, city, and state.

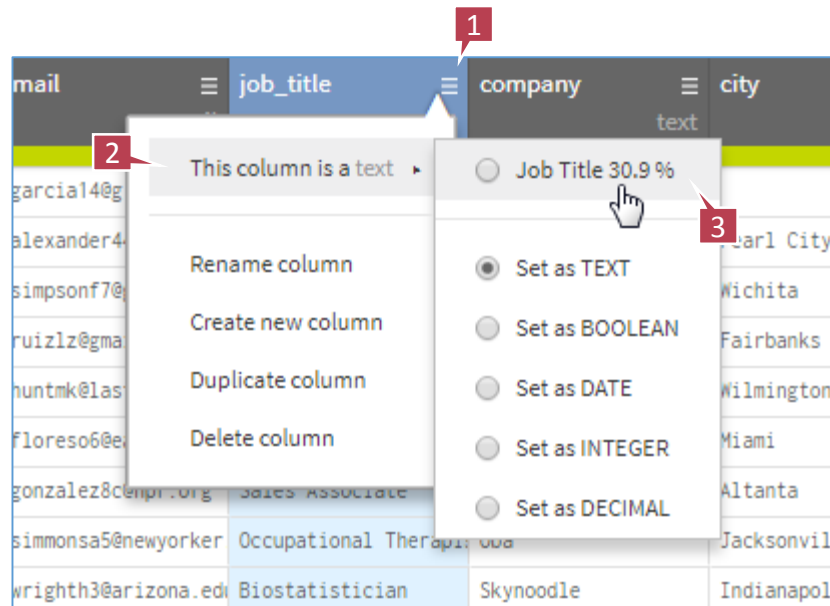
id	Name	last_name	email	job_title	company	city	state
	first name	last name	email			city	
1	Kathryn	Garcia	kgarcia14@				
2	Jason	Alexander	jalexander44@gmail.c	Chemical Engineer	Abata	Pearl City	HI
3	Lillian	Simpson	lsimpson7@gmail.com	Desktop Support Tech	Camimbo	Wichita	KS
4	Walter	Ruiz	wruiz12@gmail.com	Geological Engineer	Yakitri	Fairbanks	AK
5	Joshua	Hunt	jhuntmk@last.fm	Financial Advisor	Oyope	Wilmington	DE
6	Mildred	Flores	mflores06@earthlink.	Nurse	Edgeblab	Miami	FL
7	Victor	Gonzalez	vgonzalez8c@npr.org	Sales Associate	Ntag	Atlanta	GA
8	Joshua	Simmons	jsimmons5@newyorker	Occupational Therapi	Oba	Jacksonville	FL
9	Beverly	Wright	bwright3@arizona.ed	Biostatistician	Skynoodle	Indianapolis	IN
10	Fred	Rodriguez	frodriqueznc@fotki.c	Director of Sales	Eidel	Anchorage	AK
11	Joseph	Peterson	jpetersonn@sohu.com	Research Nurse	Gabcube	Las Vegas	NV
12	Denise	Martin	dmartin@java.com	Speech Pathologist	Zooncast	Nampa	ID
13	Jennifer	Sullivan	jsullivan4r@lycos.co	Automation Specialis	Bluezoom	Bridgeport	CT
14	Ronald	Gonzales	rgonzales5@apple.co	Automation Specialis	Shuffleitag	Racine	WI
15	Victor	Cox	vcocx9@virginia.edu	Librarian	Skalith	Bend	OR

- Multiple preparations can be saved and created for a single dataset.
- Because you created this preparation using the Add Preparation button, you do not need to save anything. Every new preparation step is automatically saved.
- Remember that the original data from your dataset remains unchanged.

Semantic Type

Semantic Type

Column heading or what the data in a specific column represents.



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. Hover 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.

Talend Data Preparation automatically suggests the proper data types for each columns of your data sets. It will help you to further discover the data. But you can change at any time those suggestions based on your own experience.

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.

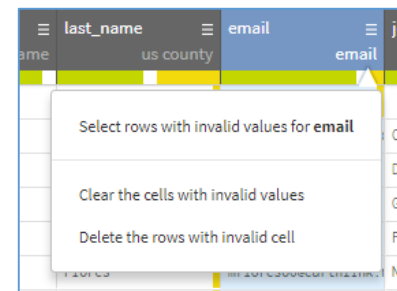
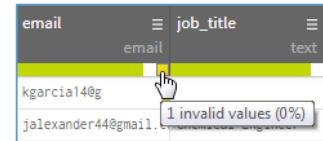
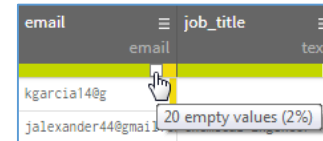
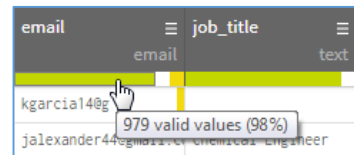
- **Green** – Data matches the cell format
- **White** – Empty cells
- **Orange** – Data in the cell does not match the cell format

id	Name	last_name	email	job_title
integer	first name	us county	email	text
<div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div>	<div><div></div><div></div><div></div></div>

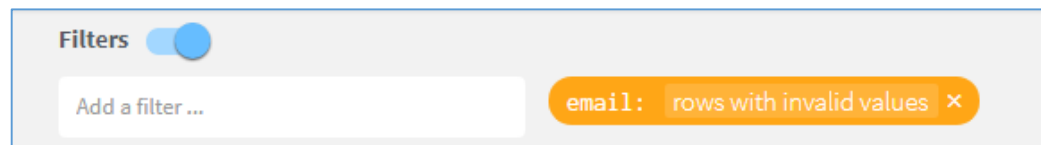
Let's take a closer look at the E-mail Data Quality Bar. Exact numbers and percentages can be found by hovering over each color.

- **Green** – 978 cells have data in the correct format
- **White** – 20 empty cells
- **Orange** – 2 cells have entries in an incorrect format

Click any color to select, delete, or clear the cells with data in an invalid format. Click the Orange section then Click Select rows with invalid values for E-Mail to display the entries with an incorrect format.



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

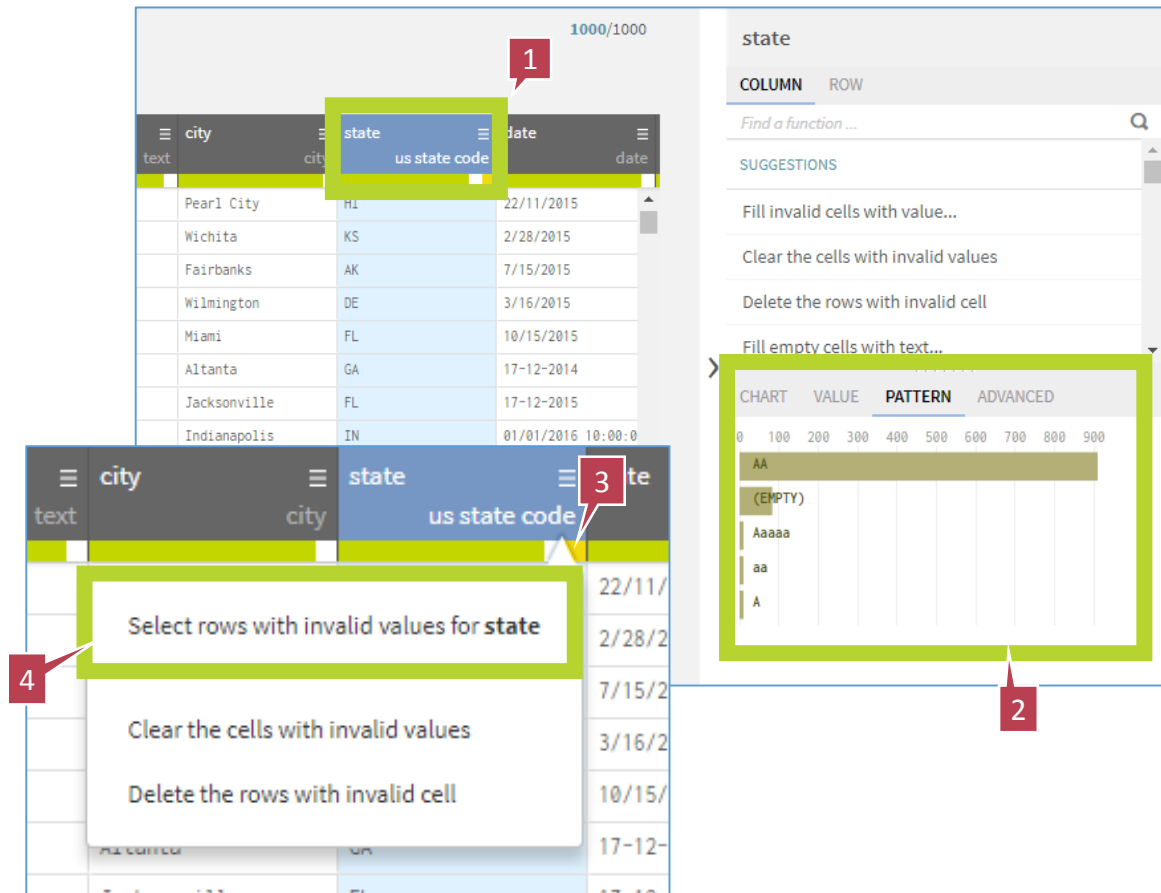


Basic Text Manipulation

To filter and fix data:

How to filter invalid rows:

1. Click the column header **STATE**
2. In the bottom right is a Pattern table. Mouse over the rows to see counts. The top row indicates that 991 records contain a 2 letter state code. **You can click a bar to view just those records (to remove the filter click the x on the box located at the top next to the filter box).**
3. On the Data Quality Bar, click the **Orange** section.
4. Click **"Selects rows with invalid values for STATE"**.
5. 7 rows will be displayed that contain invalid information.





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 data grid.

Basic Text Manipulation

To filter and fix data:

1. To edit the text value in a field **double-click in one of the cells** that contains "Texas". Change Texas to TX. **DO NOT** hit enter yet!
2. Under the cell that you are editing is a check box and the label **"Apply to all Cells 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 Code!
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 invalid records box on the Search and Filter line to return to the full list.

Filters  7/1000

Add a filter ... state: rows with invalid values 

	last_name	email	job_title	company	city	state	date	cam
	us county	email	text	text	city	us state code	date	
213	Kennedy	ekennedyg5@youtu.be	Executive Secretary	Flipopia	Cedar Rapi	TX	11/13/2015	HOCKI
754	Matthews	amattthewsg0@soup.io	Staff Scientist	Eazzy	Dallas			HOCKI
756	Lopez	jlopezio@geocities.jp	Clinical Specialist	Flipstorm	Austin			HOCKI
757	Crawford	ecrawfordjj@nasa.gov	Administrative Assis	Ozu	Dallas			BIKE
765	Shaw	jshawpm@uiuc.edu	Occupational Therapi	Roexo	Plano			TRAI
961	Webb	rwebbrk@theguardian.	Administrative Assis	Thoughtmix	Dallas	Texas	10/28/2015	HOCKI
985	Walker					E	8/11/2015	TRAI

state	date
us state code	
HI	22/11/2015
KS	2/28/2015

Recipes

1

Change to title case on column Name

2

Remove trailing and leading characters on column Name

3

Replace the cells that match on column state

state: rows with invalid values ×

Current:
= Texas

Replacement:
TX

☐ Overwrite entire cell

SUBMIT

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

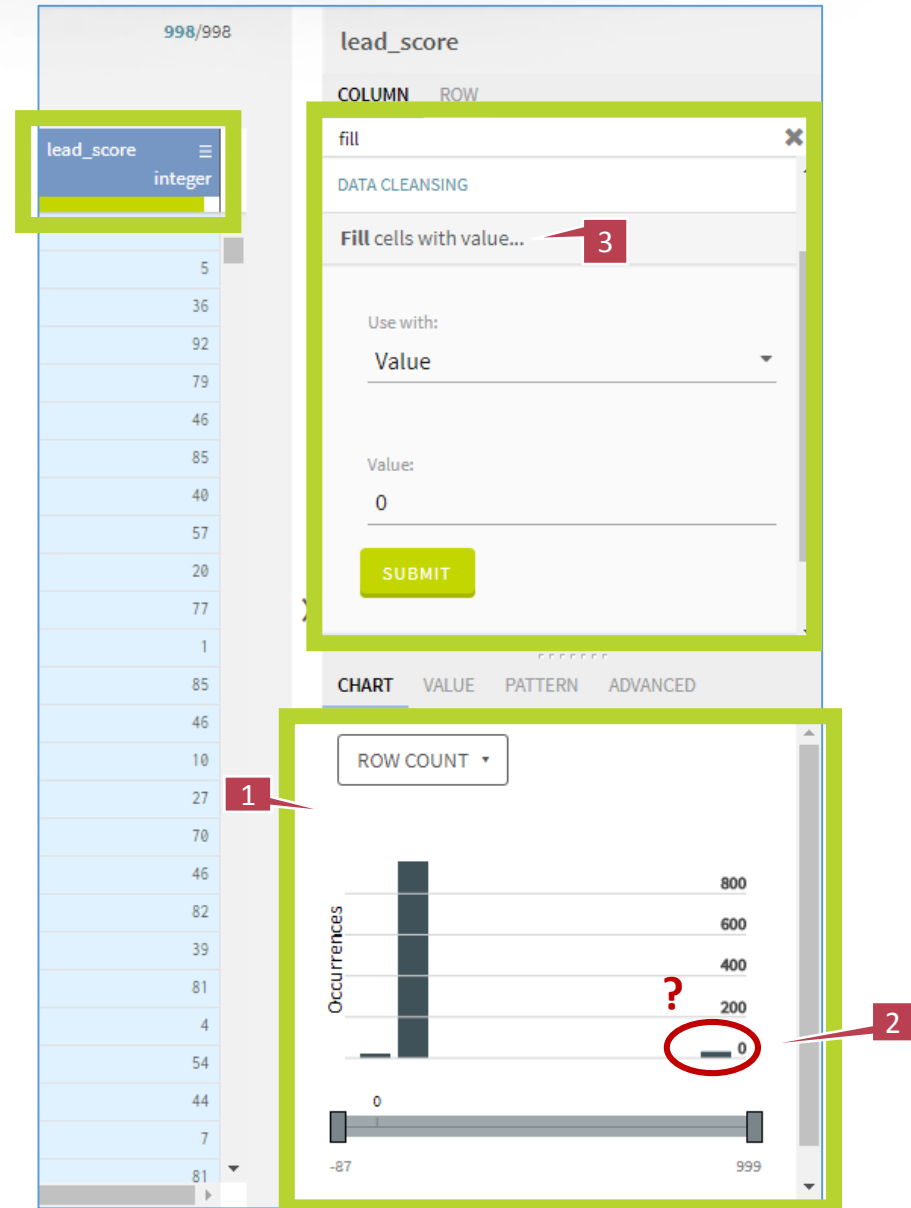
Basic Numeric Manipulation

To filter and fix data:

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

1. Click the Column **LEAD_SCORE** and you will see this is a basic integer field but look at the **Histogram Graph** at the bottom right. The data is being skewed by some larger value.
2. Click the **blue bar** on the far right on the Graph. You should 31 records with the value of 999. Looks like the Default is set to 999. This time, to change the data we will use a function called **"Fill Cell with Value ..."**.
3. Type **"Fill"** into the Search Function box 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 the Charts to help filter the data as well as change values directly in the data grid.



Basic Numeric Manipulation

To filter and fix data:

LEAD_SCORE Field, continued

1. If you take a **close look at the Graph for LEAD_SCORE column** you will notice 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 magnitude of their respective lead score numbers while dropping the negative sign.

The screenshot displays the Talend Data Studio interface. At the top, a filter is applied to the 'lead_score' column: 'lead_score in [-87..0]'. Below this, a table shows data for columns: state, date, campaign_id, and lead_score. The 'lead_score' column contains negative values: -40, -46, -5, -41, -65, -42, and -87.

On the right, a panel titled 'lead_score' shows a histogram of the data. The x-axis represents the 'lead_score' values, ranging from -87 to 105, with a maximum value of 0. The y-axis represents the 'occurrences' of each value. A red arrow points from the histogram to the 'Calculate absolute value' option in the 'SUGGESTIONS' list.

The 'SUGGESTIONS' list includes the following options:

- Delete these filtered rows
- Keep these filtered rows
- Fill empty cells with text...
- Delete the rows with empty cell
- Calculate absolute value

The 'Calculate absolute value' option is highlighted with a green box and a red arrow pointing to it.

Date Cleansing and Formatting

To filter and fix data:

Next look at the Date Field.

1. Click the **Date field**, 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 Euro Standard and others are US Standards. Some contain /'s and others dashes .
2. To standardize the dates under the **Suggested Functions** Click the **"Change Date Format"**. Select a pre-existing format or type one in. Click **Submit** once done.

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

The screenshot shows the Talend Data Preparation interface. On the left, a table with columns: state, date, campaign_id, and lead_score. The 'date' column contains various date formats. A green box highlights the 'date' column header. A red arrow points from the 'date' column to the 'SUGGESTIONS' panel on the right. In the 'SUGGESTIONS' panel, the 'date' field is selected. A green box highlights the 'Change date format...' option. A red arrow points from the 'Change date format...' option to the 'Change date format...' dialog box. The dialog box shows the current format as 'I don't know, best guess' and the new format as 'MM.dd.yyyy'. A red arrow points from the 'SUBMIT' button to the 'SUBMIT' button.

state	date	campaign_id	lead_score
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_ho1d	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	SKI_Y15Q02_vied	1
CT	6/1/2015		
WI	2.11.2015		
OR	4.6.2015		
NY	11/2/2015		

Change date format...

Current format:
I don't know, best guess

New format:
custom

Your format:
MM.dd.yyyy

SUBMIT

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 Date Format, select **custom (design your custom pattern)**. Enter **dd-MMMM-yyyy** (Date formatting is very 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 by drag & dropping**. This will save you time if you realize that a column you apply a function to, still does not fully contain expected data.

4 Fill cells with value on column
lead_score

5 Calculate absolute value on column
lead_score

6 Change date format on column date

Current format:
I don't know, best guess

New format:
custom

Your format:
MM.dd.yyyy

SUBMIT

Filters

Add a filter ...

	email	job_title	company
2	jalexander44@gmail.com	Chemical Engineer	Abata
3	lsimpsonf7@gmail.com	Desktop Support Techn	Camimbo
4	wruizlz@gmail.com	Geological Engineer	Yakitri
5	jhuntmk@last.fm	Financial Advisor	Oyope
6	mflores06@earthlink.net	Nurse	Edgeblab
7	vgonzalez8c@npr.org	Sales Associate	Ntag
8	jsimmons5@newyorker.com	Occupational Therapist	Oba
9	bwrighth3@arizona.edu	Biostatistician	Skynoodle
10	frodrigueznc@fotki.com	Director of Sales	Eidel
11	jpeterosnm@sohu.com	Research Nurse	Gabcube
12	dmartint@java.com	Speech Pathologist	Zoomcast
13	jsullivan4r@lycos.com	Automation Specialist	Bluezoom
			Shuffletag
			Skalith
			Rhyloo

Your format:
dd-MMMM-yyyy

SUBMIT

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.

The screenshot illustrates the process of masking sensitive data in Talend Data Preparation. The main window shows a data table with columns like id, Name, last_name, email, title, company, city, state, date, campaign_id, and lead_score. The 'email' column is selected, and the 'Mask data (Obfuscation)' function is applied. A pop-up window shows the function's settings, and another window shows the resulting masked data where the part of the email before the '@' symbol is replaced by 'X's.

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.

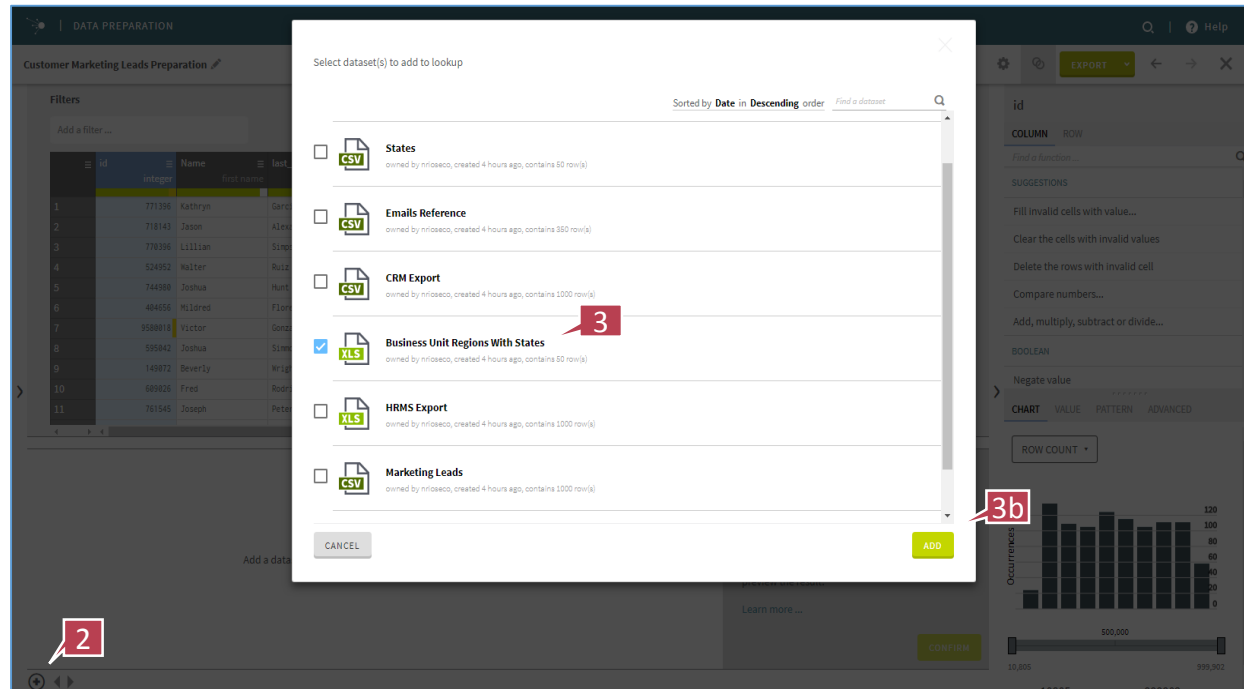
When manipulating sensitive data such as names, addresses, credit card or social security numbers, you might want to mask this data. To protect the original data, you will use the Data Masking function to generate functional substitutes.

Data Blending

To blend data:

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 files are available to choose from by clicking the **+** icon.
3. Click the **Checkbox** in front of “Business Unit Regions with States” and then click **Add**.



Data Blending

Data Blending Continued:

1. Click the **column you would like to blend**, which is the **state** column in your current dataset.
2. At the bottom, have the Region added by clicking **Add to Dataset** under the Region column header.
3. **Hover over Confirm** to preview the changes which display in yellow. To accept the changes, **click Confirm**.

The screenshot shows the Talend Data Preparation interface for a dataset named 'Customer Marketing Leads Preparation'. The main table displays customer data with columns: first_name, last_name, email, job_title, company, city, state (us_state_code), Region, date, and campaign_id. A red callout '1' points to the 'state' column header. Below the main table, a 'State' dropdown menu is open, showing a list of US states. A red callout '2' points to the 'Add to Dataset' button in the 'Region' column header. A red callout '3' points to the 'CONFIRM' button in the bottom right corner of the interface. The interface also includes a 'Filters' section, a 'SUGGESTIONS' panel on the right, and a 'CHART' section at the bottom right.

	first_name	last_name	email	job_title	company	city	state	Region	date	campaign_id
2	Alexander	XXXXXXX@gmail.c	Chemical Engineer	Abata	Pearl City	HI	West	11.22.2015	HOOKEY_Y15Q1_cant	
3	Simpson	XXXXXXXXX@gmail.com	Desktop Support Techn	Cambo	Wichita	KS	Mid West	02.28.2015	RUN_Y14Q2_deal	
4	Ruiz	XXXXXXXXX@gmail.com	Geological Engineer	Yaktr	Fairbanks	AK	West	07.15.2015	TRAIL_Y14Q4_purr	
5	Hunt	XXXXXXXXX@ast.fm	Financial Advisor	Oyope	Wilmington	DE	North East	03.16.2015	HOOKEY_Y14Q2_node	
6	Flores	XXXXXXXXX@earthlink.i	Nurse	EdgeLab	Miami	FL	South East	10.15.2015	HOOKEY_Y15Q4_chun	
7	Gonzalez	XXXXXXXXX@npr.org	Sales Associate	Ntag	Atlanta	GA	South East	12.17.2014	TRAIL_Y15Q3_hoid	
8	Simmons	XXXXXXXXX@newyorker	Occupational Therapi	Oba	Jacksonville	FL	South East	12.17.2015	TRAIL_Y14Q3_moon	
9	Wright	XXXXXXXXX@arizona.ed	Biostatistician	Skyoodle	Indianapolis	IN	Mid West	01.01.2016	TRAIL_Y15Q4_ross	
10	Rodriguez	XXXXXXXXX@fotki.c	Director of Sales	Eisel	Anchorage	AK	West	07.06.2015	BIKE_Y14Q2_hurt	
11	Peterson	XXXXXXXXX@sohu.com	Research Nurse	Gabcube	Las Vegas	NV	West	03.16.2015	HOOKEY_Y15Q2_boos	
12	Martin	XXXXXXXXX@java.com	Speech Pathologist	Zoomcast	Nampa	ID	West	12.09.2014	SKL_Y15Q2_vied	
13	Sullivan	XXXXXXXXX@lycos.co	Automation Specialis	Bluzoon	Bridgeport	CT	North East	06.01.2015	SKL_Y14Q3_vack	
14	Gonzales	XXXXXXXXX@apple.co	Automation Specialis	Shuffletag	Racine	WI	Mid West	11.02.2015	HOOKEY_Y14Q4_roan	
15	Cox	XXXXXXXXX@virginia.edu	Librarian	Skalth	Bend	OR	West	06.04.2015	TRAIL_Y15Q4_hays	
16	Wilson	XXXXXXXXX@va.gov	Actuary	Rhyloo	Manhattan	NY	North East	11.02.2015	TRAIL_Y14Q4_fete	
17	Arnold	XXXXXXXXX@youtube.co	Senior Editor	Tazzy	Columbus	GA	South East	12.25.2014	HOOKEY_Y15Q1_file	

State: us state code

Region: Add to Dataset

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 ...

CONFIRM

Group and Standardize

To Group data:

Group and Standardize allows you to find cells that have similar text 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 **Find and Group Similar Text**.

The screenshot displays the Talend Group and Standardize tool interface. The main table on the left lists job titles, companies, cities, and states. The right panel shows the 'job_title' column selected, with the search field containing 'group'. The 'Find and Group Similar Text' button is highlighted. Below the search field, a horizontal bar chart shows the frequency of job titles, with 'Occupational Therapist' being the most frequent.

job_title	company	city	state
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	Atlanta	GA
Occupational Therapist	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 Specialist	Bluezoom	Bridgeport	CT
Automation Specialist	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 Associate	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 Administrator	Fivebridge	Greenville	DE
Junior Executive	Kwimbee	Wilmington	DE
Librarian	Feedfish	Pike Creek	DE

The right panel shows the 'job_title' column selected, with the search field containing 'group'. The 'Find and Group Similar Text' button is highlighted. Below the search field, a horizontal bar chart shows the frequency of job titles, with 'Occupational Therapist' being the most frequent.

Find and Group similar text

To group data:

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 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 **uncheck** the box in front of the job title.
4. If you do not want to change a group of job titles **uncheck** the box in the first column.
5. Click **Submit** when finished.

FIND AND GROUP SIMILAR TEXT

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

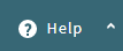
<input checked="" type="checkbox"/>	These values have been found	1	This value will be kept
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Health Coach <input checked="" type="checkbox"/> Health Coach!	3	Replace value: Health Coach
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Administrative Assistant <input checked="" type="checkbox"/> Administrative Officer		Replace value: Administrative Assistant
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Account Executive <input checked="" type="checkbox"/> Account Representative <input checked="" type="checkbox"/> Account Representativel <input checked="" type="checkbox"/> Accountant <input checked="" type="checkbox"/> Accounting Assistant	4	Replace value: Accountant

SUBMIT

5

Feedback

Your feedback is very important to us. We want to know whether our services are meeting your needs and being delivered effectively.

To leave feedback, click the arrow next to  and select **Feedback** from the drop-down list. Fill out the form with e-mail and message. See also the links to the forums and knowledge base in this feedback form

×

Send feedback

Enter your email

Summary (required)

Bug

▼

Minor

▼

Description

Feel free to ask questions and interact with us [on the forum](#)

Check out our documentation, knowledge base, videos etc. [online](#)

CANCEL

OK

Conclusion



Now, your data is ready:

- **For analysis:** for example, you can analyze your leads score by date or state into Excel or Tableau now that your data is cleansed and standardized and once you have exported your preparation results.
- **For further integration:** the data has been cleansed and formatted so that it is ready to be uploaded into CRM or Marketing Automation application such as Marketo or Salesforce.com.

The good news is that ...

With Talend, Data is just one click away from everyone's daily task.

What are your next steps?

Now you ready to add your own datasets, run your own data preparations... and turn your daily tasks into data-driven activities.

