



Talend Data Preparation Free Desktop

Guide de démarrage
V2.1





Formation à Talend Data Preparation

Guide d'introduction

Pour aller à une section spécifique du guide, cliquez sur les menus ci-dessous :

**Présentation
de Talend
Data
Preparation**

**Accès et
démarrage de
Talend Data
Preparation**

**Exercices
faciles de
nettoyage de
données**

**Opérations
basiques de
manipulation
de données**

**Nettoyage et
formatage de
dates**

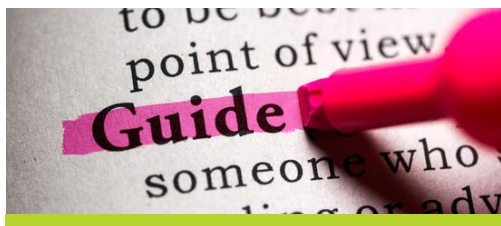
Formation à Talend Data Preparation

Guide d'introduction

[Présentation de Talend Data Preparation](#)[Accès et démarrage de Talend Data Preparation](#)[Exercices faciles de nettoyage de données](#)[Opérations basiques de manipulation de données](#)[Nettoyage et formatage de dates](#)

À propos de ce guide

Qu'est-ce que le guide d'introduction de Talend Data Preparation ?



Le guide d'introduction de Talend Data Preparation fournit des instructions étape par étape qui vous permettront de créer et de réaliser de A à Z des scripts de préparation de données.



La démo a été conçue à partir de cas d'utilisation réels avec des données marketing. Ils reflètent les problématiques auxquelles de nombreux utilisateurs sont confrontés tous les jours, sur des tableaux les obligeant à utiliser des macros complexes ou, pis, des scripts VBA.



Le but est de vous familiariser avec cet outil aussi intuitif qu'une interface utilisateur sur le web. Vous comprendrez comment Talend peut vous aider à découvrir, nettoyer, mettre en forme et enrichir vos données.

À propos de Talend Data Preparation

L'outil de préparation de données en libre-service pour tous



Regardez la vidéo
[Introduction to Talend Data Preparation Video](#)



Des données nettoyées et utiles en seulement quelques minutes (au lieu de plusieurs heures)

- Un point d'entrée unique pour tout type de source de données
- Découverte, nettoyage et formatage interactifs
- Automatisez et réutilisez les formules de préparation de données



L'outil de nettoyage de données en libre-service pour tous

- Mettez les données à votre service dans vos tâches quotidiennes
- Découvrez et explorez vos données
- Laissez-vous guider sur votre chemin vers des données actionnables.



Donnez de l'autonomie aux tâches informatiques et profitez plus rapidement de connaissances approfondies

- Accès aux données maîtrisé et en libre-service pour tous
- Évitez des incohérences et des fuites de données nuisibles à l'entreprise
- Soulagez les équipes informatiques et dynamisez la productivité

Si vous avez déjà installé Talend Data Preparation, [cliquez ici](#) pour passer les instructions d'installation.

Marketing Lead Data Preparation

Les données du fichier « customer marketing leads.csv » concernent des leads. Elles comportent des problèmes de qualité, et de nombreux champs doivent donc être reformattés. Analyser les données brutes dans ce fichier aboutirait à des résultats décevants, à cause d'informations incorrectes ou de valeurs manquantes dont la correction avec Excel prendrait des heures.

Dans cette démo, nous vous accompagnerons dans des actions faciles de préparation de données qui, sur Excel, vous donneraient du fil à retordre.

Vous découvrirez comment :

- Changer rapidement les valeurs des données après les avoir identifiées grâce à des graphiques et à des filtres très simples, sans codage !
- Tester des éléments formidables tels que les histogrammes pour corriger les données !
- Manipuler du texte, des dates et des champs numériques dans le même fichier en seulement quelques clics !

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

Prérequis pour Talend Data Preparation

Voici les informations à propos des logiciels et du matériel recommandés pour commencer avec Talend Data Preparation :

Prérequis matériel :

Processeur	Processeur 64 bits requis (Note: les processeurs 32 bits ne sont pas supportés)
Mémoire allouée	Minimum 1 Go
Espace sur le disque	Minimum 500 Mo

Prérequis logiciels :

Système d'exploitation	<ul style="list-style-type: none">Windows 7 ou plus récentMac OS X 10.7 « Lion » ou plus récent
------------------------	--

Navigateurs web compatibles :

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

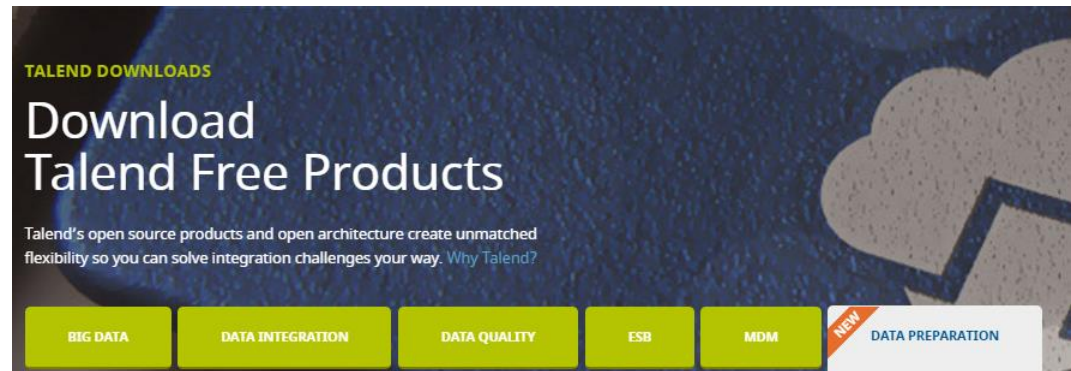
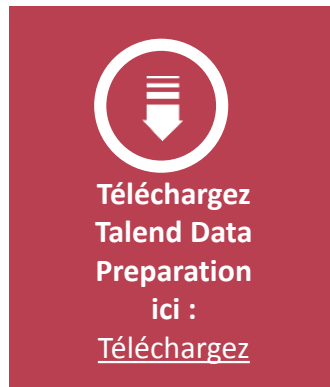
Java :

Il n'y a pas de besoin spécifique pour la plupart des ordinateurs Windows et Apple. Cependant, si vous souhaitez installer la version Apache de Talend Data Preparation, vous devez avoir Oracle Java 8 (64 bits) installé sur votre ordinateur. La version par défaut pour Windows 32 bits n'est pas supportée, seule la version 64 bits est supportée.

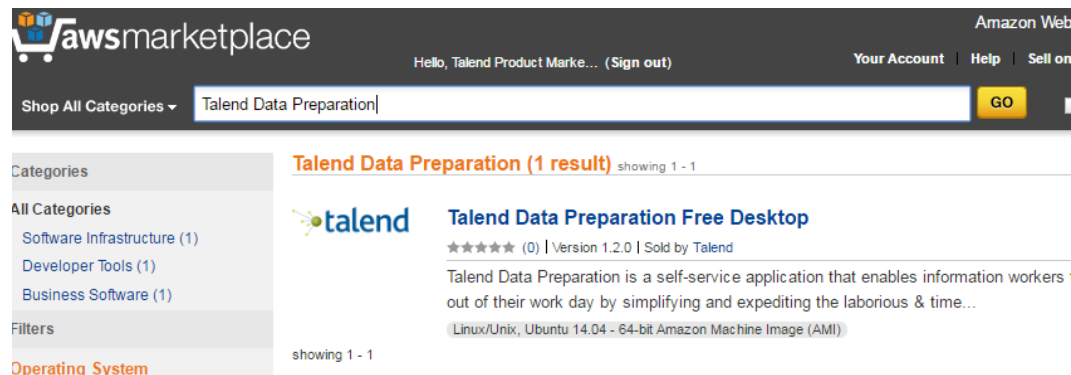
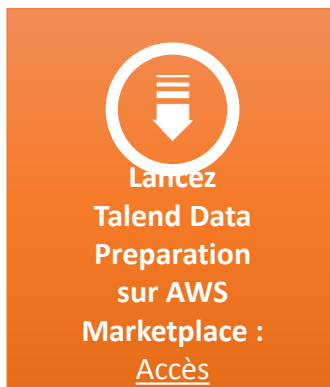
Comment accéder à Talend Data Preparation ?

Vous pouvez accéder à Talend Data Preparation :

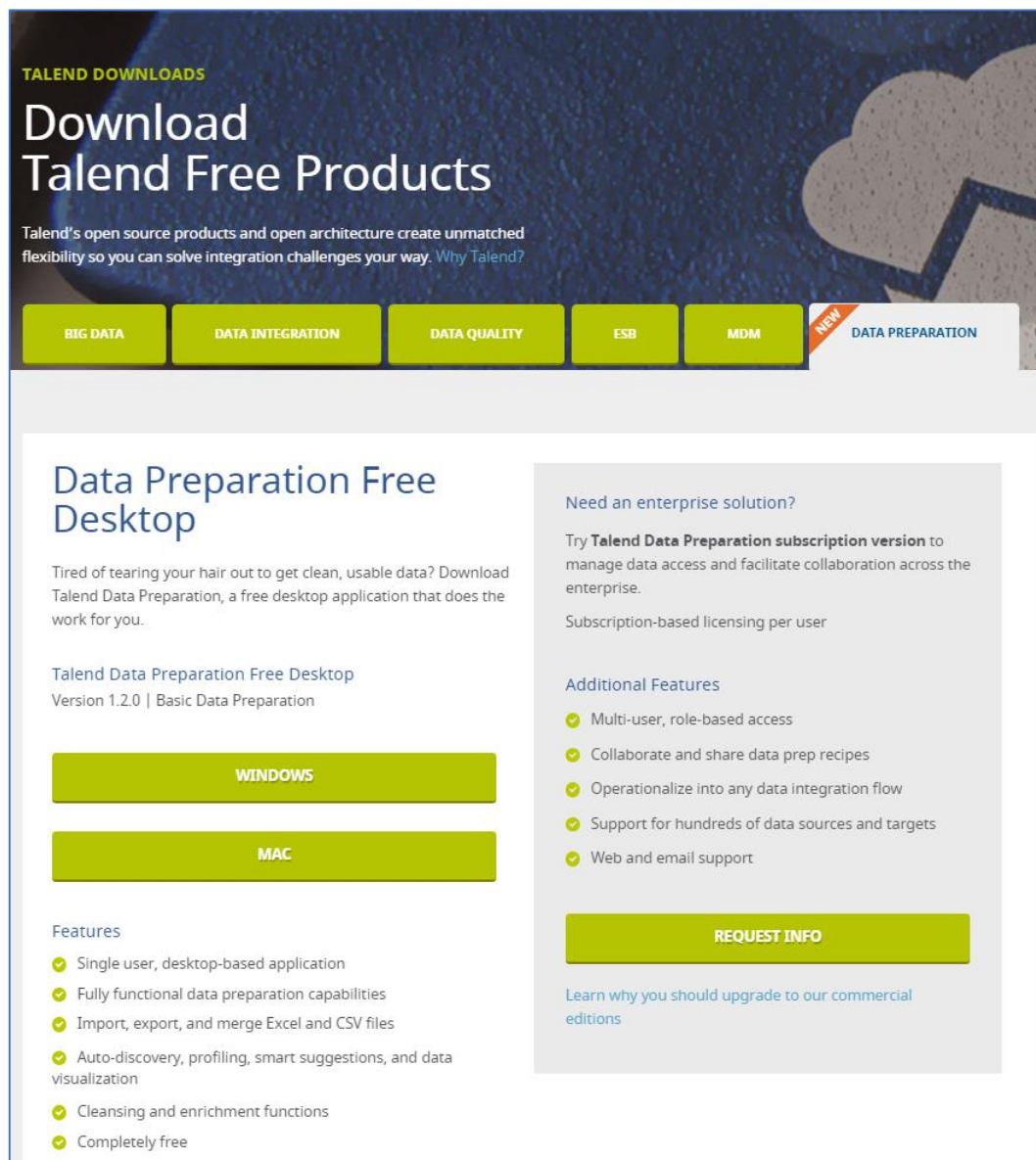
- Soit en téléchargeant directement [sur votre ordinateur](#)



- Soit [à La Demande en Cloud](#) en lançant Talend Data Preparation sur une machine virtuelle Amazon (AMI) depuis l'AWS Marketplace



Comment télécharger Talend Data Preparation ?



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](#)



Téléchargez
Talend Data
Preparation
ici :
[Téléchargez](#)

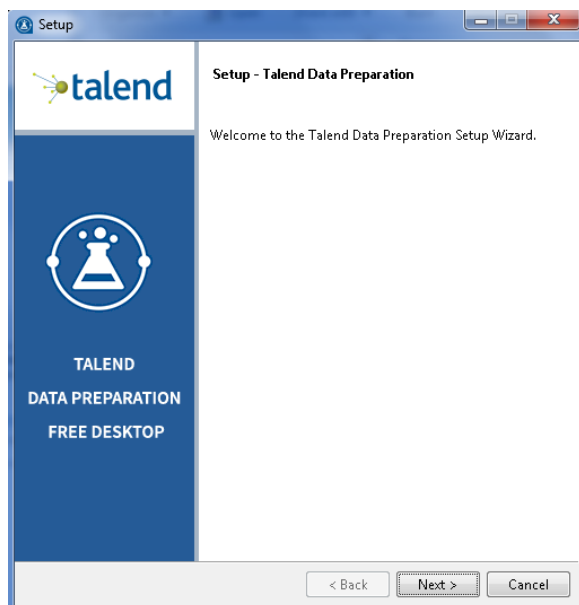
Comment installer Talend Data Preparation sur Windows ?



La version pour Windows est fournie en tant qu'installateur Microsoft Windows standard. Vous aurez besoin des droits administrateur pour l'exécuter. Pour installer et démarrer Talend Data Preparation veuillez suivre les étapes suivantes :

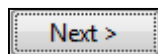
1

Après avoir téléchargé le fichier (voir page précédente), double-cliquez sur **Talend-DataPreparation-Free-Desktop-2.1.exe**



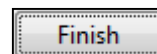
2

Cliquez sur **Next** pendant l'installation et utilisez les paramètres par défaut



3

Cliquez sur **Finish** une fois l'installation terminée



4

Pour commencer à utiliser Talend Data Preparation, cliquez sur l'icône du programme ou sur le raccourci dans le menu Démarrer.



Installation alternative pour les utilisateurs Windows



Si vous n'avez pas les droits administrateurs requis pour utiliser l'installateur, suivez les étapes ci-dessous pour installer Data Preparation via un fichier .zip :

1

Sur la page de téléchargement de Talend Data Preparation, descendez jusqu'à la section **Other Releases**.

2

Téléchargez le fichier **Talend-DataPreparation-Free-Desktop-windows-2.1.0.zip**

3

Dézippez le fichier où vous le souhaitez sur votre ordinateur.

4

Lancez le fichier **.exe** pour utiliser l'outil Talend Data Preparation

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

Comment installer Talend Data Preparation sur Mac OS X ?



Pour installer et démarrer Talend Data Preparation veuillez suivre les étapes suivantes :

1

Double-cliquez sur le fichier Talend-DataPreparation-Free-Desktop-2.1.dmg pour ouvrir le dossier.

2

Glissez et déplacez l'icône Talend dans le dossier Applications.

3

Talend Data Preparation figure maintenant dans la liste de vos applications. Ouvrez-le avec un double-clic.

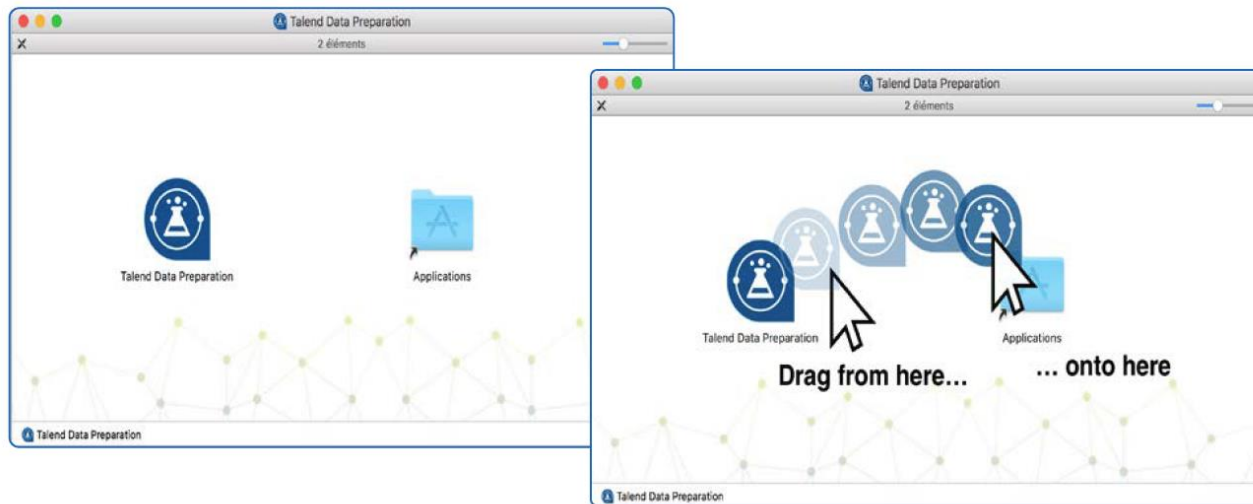
4

Pour désactiver App Nap et assurer des performances optimales, suivez cette procédure rapide :

1. Ouvrez le Terminal à partir du dossier
/Applications/Utilities.

2. Exécutez la commande suivante :

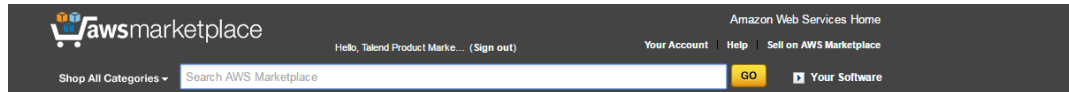
```
defaults write  
org.talend.dataprep  
NSAppSleepDisabled -bool  
YES
```



Comment accéder à Talend Data Preparation ?



Lancez
Talend Data
Preparation
sur AWS
Marketplace :
Accès



Talend Data Preparation Free Desktop

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

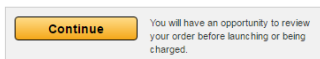
Talend Data Preparation is a self-service application that enables information workers to cut hours out of their work day by simplifying and expediting the laborious & time consuming process of preparing data for analysis or other data-driven tasks. It enables anyone to quickly discover, cleanse, standardize, 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-gsSelf-service data cleansing for anyoneEmpower 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.



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

Pour **lancer Talend Data Preparation sur AWS Marketplace** :

- Connectez-vous avec votre compte AWS ou créez-en un
- Confirmez votre abonnement : le logiciel est gratuit, vous ne serez facturé que des frais d'utilisation d'infrastructure par AWS
- Configurez et lancez votre instance AMI
- Connectez votre instance
- Lancez le logiciel

Page principale – Préparations et jeux de données

Une fois l'application démarrée, la première page qui apparaît à l'écran est la page « Preparations » :

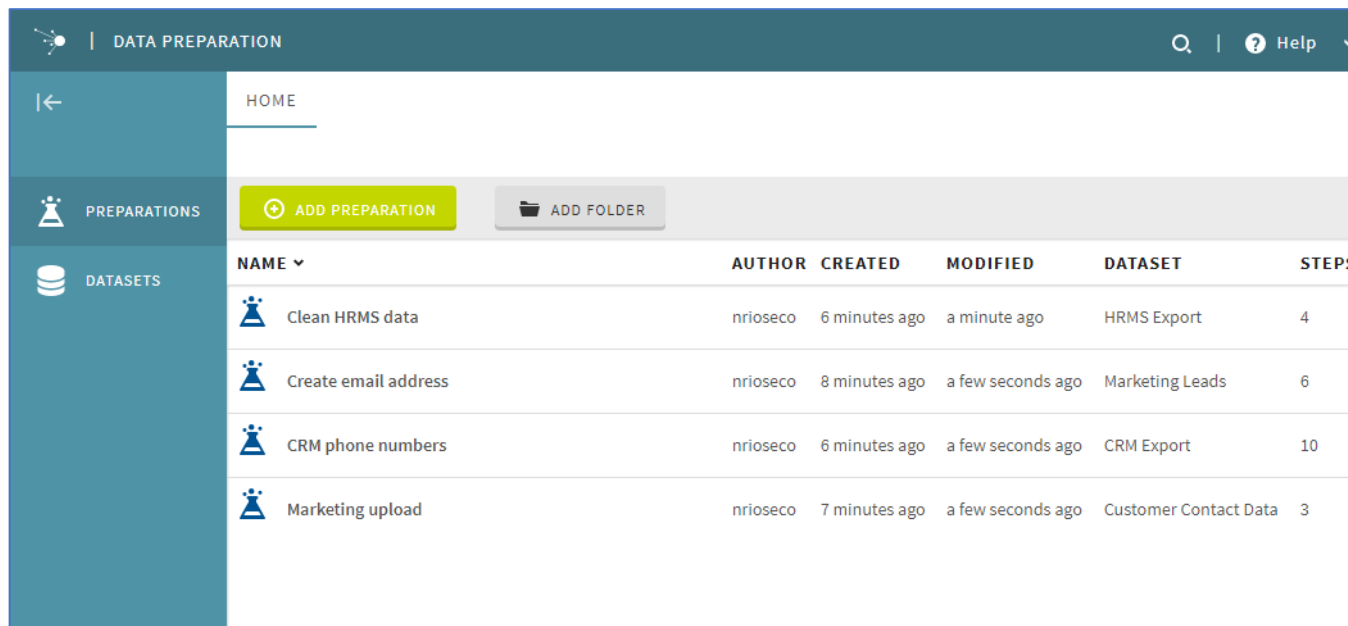
Dans la liste, vous verrez toutes les préparations sur lesquelles vous avez travaillé. Une préparation est le résultat des différentes étapes que vous avez appliqué pour nettoyer vos données. Vous pouvez exporter ce résultat en tant que fichier. Une préparation prend un jeu de données en entrée et applique la recette pour produire le résultat final. Les données d'origine ne sont jamais modifiées.

Depuis cette page, vous pouvez aussi accéder à la vue « Datasets » :

Vous verrez ici tous les jeux de données sur lesquels vous avez travaillé ou que vous avez importé. Les jeux de données peuvent être des fichiers en local ou à distance pouvant être importés dans Talend Data Preparation. Dans la version commerciale de Talend Data Preparation, ils peuvent également provenir d'une connexion à une base de données ou d'autres sources de données. Les jeux de données sont utilisés comme matériaux de base d'une ou plusieurs préparations.

À partir de cette page vous pouvez :

1. Ajouter de nouvelles préparations
2. Organiser vos préparations en dossiers
3. Importer et créer de nouveaux jeux de données
4. Enregistrer vos jeux de données en tant que favoris

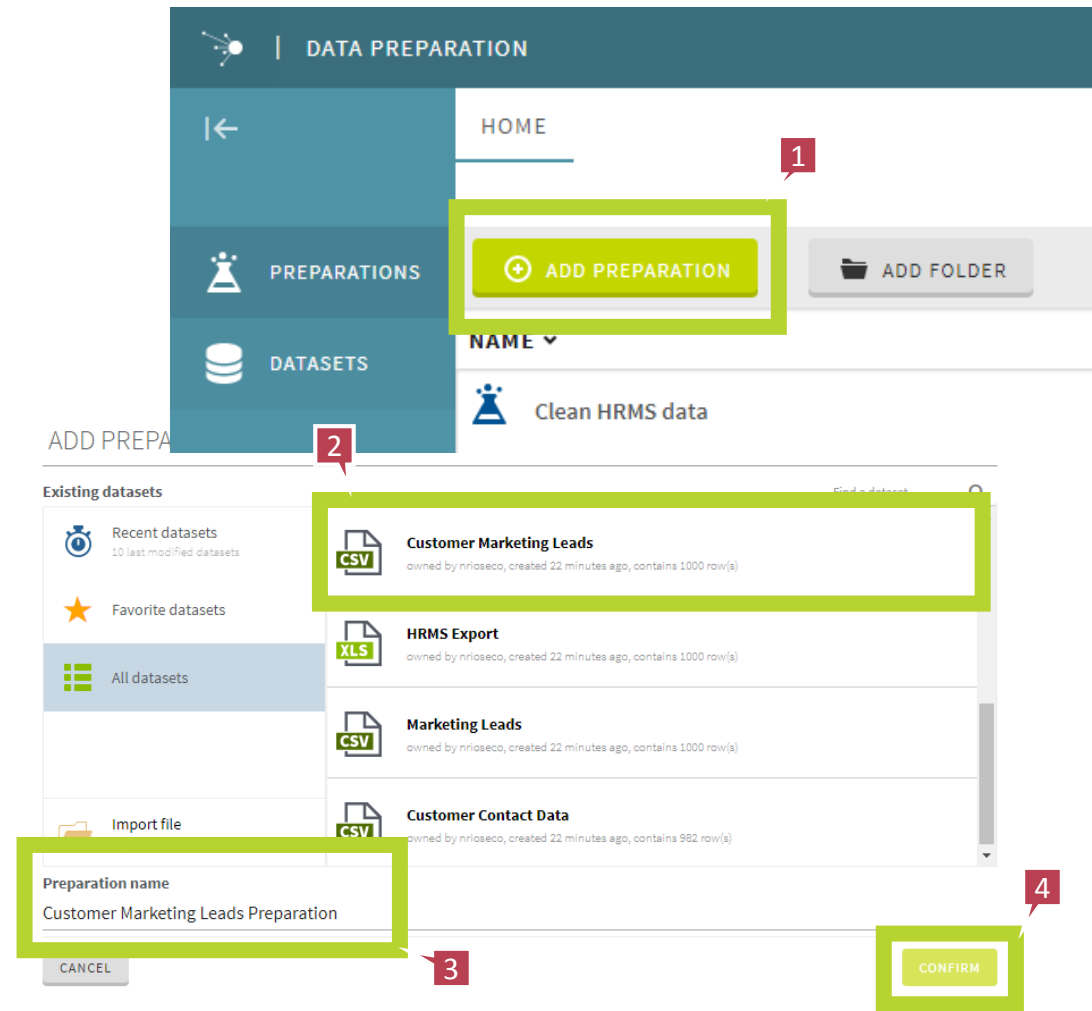


DATA PREPARATION		Q ? Help ▾				
←	HOME					
PREPARATIONS	ADD PREPARATION	ADD FOLDER				
DATASETS	NAME ▾	AUTHOR	CREATED	MODIFIED	DATASET	STEPS
	Clean HRMS data	nrioseco	6 minutes ago	a minute ago	HRMS Export	4
	Create email address	nrioseco	8 minutes ago	a few seconds ago	Marketing Leads	6
	CRM phone 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

Comment ajouter une préparation ?

Pour commencer avec notre exemple :

1. Cliquez sur le bouton **Add Preparation** dans la vue « Preparation ».
2. La fenêtre **Add Preparation** s'ouvre. Cliquez sur le jeu de données **Customer Marketing Leads** dans la liste « All Datasets ».
3. Choisissez un **nom** pour votre préparation.
4. Cliquez sur **Confirm** pour ouvrir la préparation et commencer à nettoyer les données du jeu de données.



Visite guidée de Data Preparation

Dans cette **démo**, nous vous montrerons des...

**Exercices faciles de
nettoyage de
données**

**Manipulations de
données**

**Fonctions de
standardisation et
d'enrichissement
avancées**

Exemples de nettoyage simples

Effectuer des opérations basiques de formatage et de nettoyage

Occupons-nous d'abord de la colonne
« Name ».

1. Déplacez-vous sur la colonne
« NAME » et cliquez sur l'en-tête de
la colonne.
2. Le menu situé en haut à droite de
l'écran affiche la liste des fonctions
disponibles. Afin de corriger les
données, vous pouvez choisir celle qui
vous convient ou bien utiliser la
fonction suggérée.
3. Selon la résolution de votre écran,
vous devrez peut-être descendre dans
la liste pour trouver la fonction qui
permet de convertir les lettres en
majuscules au début du prénom,
« Change Style to Title Case ». En
restant sur le bouton sans cliquer,
vous pourrez obtenir un aperçu des
modifications. Cliquez pour
appliquer les changements.

The screenshot shows the Talend Data Preparation interface with a table of customer data. The 'Name' column is highlighted, and the 'Functions' menu is open, showing various data manipulation options. The table data is as follows:

id	first name	last name	email	job_title	company	city	state	date	campaign_id	lead_score
1	Kathryn	Agarwal	agarcwa4@gmail.com	Chemical Engineer	Abata	Pearl City	KS	22/11/2015	HOOKEY_Y15Q01_cnt	5
2	Jason	Alexander	jalexander44@gmail.com	Desktop Support Tech	Camsbo	Wichita	KS	2/28/2015	RUL_Y14Q02_deal	36
3	Lillian	Simpson	lsimpson7@gmail.com	Geological Engineer	Yakistri	Fairbanks	AK	7/15/2015	TRAIL_Y14Q04_purr	92
4	WALTER	Ruiz	wruiz2@gmail.com	Financial Advisor	Oyope	Wilmington	DE	3/16/2015	HOOKEY_Y14Q02_node	79
5	Joshua	Hunt	jhunt8@iact.fr	Nurse	Edgeblab	Miami	FL	10/15/2015	HOOKEY_Y15Q04_chum	46
6	Mildred	Flores	mflores0@earthlink.net	Sales Associate	Wtag	Atlanta	GA	17-12-2014	TRAIL_Y15Q03_hold	85
7	Victor	Gonzalez	vgonzalez0@npr.org	Occupational Therapist	Oba	Jacksonville	FL	17-12-2015	TRAIL_Y14Q03_moon	40
8	Joshua	Simons	jsimonsa@newyorker.com	Biostatistician	Skynoodie	Indianapolis	IN	01/01/2016 10:00:00	TRAIL_Y15Q04_rossy	57
9	Beverly	Wright	bwright3@marizona.edu	Research Nurse	Gabcube	Anchorage	AK	7/6/2015	BXKE_Y14Q02_hurt	20
10	Fred	Rodriguez	frrodriguez0@fotki.com	Speech Pathologist	Zoomcast	Nampa	ID	12/9/2014	SKL_Y15Q02_vied	1
11	Denise	Peterson	dpeterson@nahu.com	Automation Specialist	Bluezoom	Bridgeport	CT	6/1/2015	SKL_Y14Q03_jack	85
12	Martin	Sullivan	msullivan4@ycom.com	Automation Specialist	Shuffletag	Racine	WI	2.11.2015	HOOKEY_Y14Q04_roan	46
13	Jennifer	Gonzales	jgonzales5@apple.com	Librarian	Skailth	Bend	OR	4.6.2015	TRAIL_Y15Q04_hays	10
14	Ronald	Cox	vcoc9@virginia.edu					11/2/2015	TRAIL_Y14Q04_fete	27
15	VICTOR							12/25/2014	HOOKEY_Y15Q01_file	70

The 'Functions' menu is open, showing the 'Name' column selected. The 'Functions' list includes:

- Find a function...
- SUGGESTIONS
- Fill empty cells with text...
- Delete the rows with empty cell
- Mask data (obfuscation)
- Change to upper case
- Change to lower case
- BOOLEAN
- Negate value
- COLUMNS
- Concatenate with...

The 'Name' column header is highlighted, and the 'first name' sub-header is also highlighted. The 'Functions' menu is open, showing the 'Name' column selected. The 'Functions' list includes:

- Find a function...
- STRINGS
- Calculate length
- Change to lower case
- Change to title case
- Change to upper case
- Contains text...
- Extract parts of the text...

Dans cette étape nous nettoyons les champs contenant les prénoms des clients afin d'effectuer une standardisation de base. Comme vous pouvez le constater, les prénoms commencent soit avec une minuscule, soit avec une majuscule. Les espaces redondants et les noms de famille sont reconnus comme formats incorrects.

Exemples simples de nettoyage

Effectuer des opérations basiques de formatage et de nettoyage

Colonne "Name" (suite)

1. En observant les données, vous constaterez que des petits carrés blancs sont affichés avant ou après certains prénoms, comme par exemple « Joshua ».
2. Pour supprimer ces carrés blancs, cherchez puis sélectionnez la fonction appropriée, « **Remove Whitespaces (trailing and Leading)** » et cliquez sur « **Submit** ».

The screenshot shows the Talend Data Preparation interface with a table titled 'Customer Marketing Leads Preparation'. The table has columns: id, Name, last_name, email, job_title, company, city, state, date, campaign_id, and lead_score. The 'Name' column contains entries like 'Joshua', 'Beverly', 'Fred', and 'Joseph'. A red arrow points to the 'Joshua' entry. A green box highlights the 'Name' column header. A second green box highlights the 'Remove trailing and leading characters...' suggestion in the 'SUGGESTIONS' panel. A red arrow points from the suggestion to the 'Joshua' cell.

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	wrutz2@gmail.com	Financial Advisor	Oyope	Wilmington	DE	3/16/2015	HOOKEY_Y14002_node	79
5	744980 Joshua	Hunt	jhunt2@gmail.com	Nurse	Edgelab	Miami	FL	10/15/2015	HOOKEY_Y15004_shum	46
6	404656 Mildred	Flores	mflores@gmail.com	Sales Associate	Wag	Atlanta	GA	17-12-2014	TRAIL_Y15003_hold	85
7	958018 Victor	Gonzalez	vgonzalez@gmail.com	Occupational Therapist	Oha	Jacksonville	FL	17-12-2015	TRAIL_Y14003_moon	40
8	149072 Beverly	Wright	bwright3@gmail.com	Biostatistician	Sky noodle	Indianapolis	IN	01/01/2016 10:00:00	TRAIL_Y15004_ross	57
9	609026 Fred	Rodriguez	frrodriguez@gmail.com	Director of Sales	Eidel	Anchorage	AK	7/6/2015	BEKE_Y14002_hurt	20
10	761545 Joseph	Peterson	jpeterson@gmail.com	Research Nurse	Gabcube	Las Vegas	NV	3/16/2015	HOOKEY_Y15002_boos	77
11	31599 Denis	Martin	dmartin@gmail.com	Speech Pathologist	Zoomcast	Nampa	ID	12/9/2014	SKL_Y15002_vied	1
12	955467 Dennis	Sullivan	jsullivan@gmail.com	Automation Specialist	Bluezoom	Bridgeport	CT	6/1/2015	SKL_Y14003_yack	85
13	43871 Ronald	Gonzales	rgonzales@gmail.com	Shufflatag	Racine	Racine	WI	2-11-2015	HOOKEY_Y14004_roam	46
14	388630 VICTOR	Cox	vcocox@gmail.com	Librarian	Skalith	Bend	OR	4-6-2015	TRAIL_Y15004_hays	10
15	698310 Cathie	Wilson	cwilson@gmail.com	Actuary	Rhyloo	Manhattan	NV	11/22/2015	TRAIL_Y14004_fete	27
16	542272 Andy	Arnold	aarnold@gmail.com	Senior Editor	Tazzy	Columbus	GA	12/25/2014	HOOKEY_Y15003_fille	70
17	576157 Kenna	Harper	kharpers@gmail.com	Structural Engineer	Dynava	Overland Park	KS	8/31/2015	HOOKEY_Y14003_tone	46
18	217977 Bruce	Richards	brichards@gmail.com	Help Desk Operator	Gabtube	Orange	CT	4/4/2015	HOOKEY_Y15004_mine	82
19	874029 Brando	Porter	bporter@gmail.com	Senior Sales Associate	Nath	Cheshire	CT	12/22/2014	HOOKEY_Y15001_tute	39
20	133382 Angus	Stone	astone@gmail.com	VP Marketing	Oozz	New Haven	CT	5/31/2015	HOOKEY_Y15002_eggs	81

Recettes

1. Chaque fois que vous sélectionnez une fonction, elle s'ajoute automatiquement à la « recette » située dans le panneau de gauche.
2. **Pour supprimer un élément de la recette**, placez le curseur sur la ligne correspondante et cliquez sur la corbeille.
3. **Pour renommer une préparation**, cliquez sur l'icône en forme de crayon et entrez un nouveau nom.
4. La recette peut être masquée en cliquant sur la flèche.
5. **Pour exporter** le résultat de votre préparation, cliquez sur « export » puis sélectionnez le type de fichier souhaité.

Customer Marketing Leads Preparation

1 Change-to-title-case-on-column-Name

2 Remove step

3 DATA PREPARATION

4

5 EXPORT

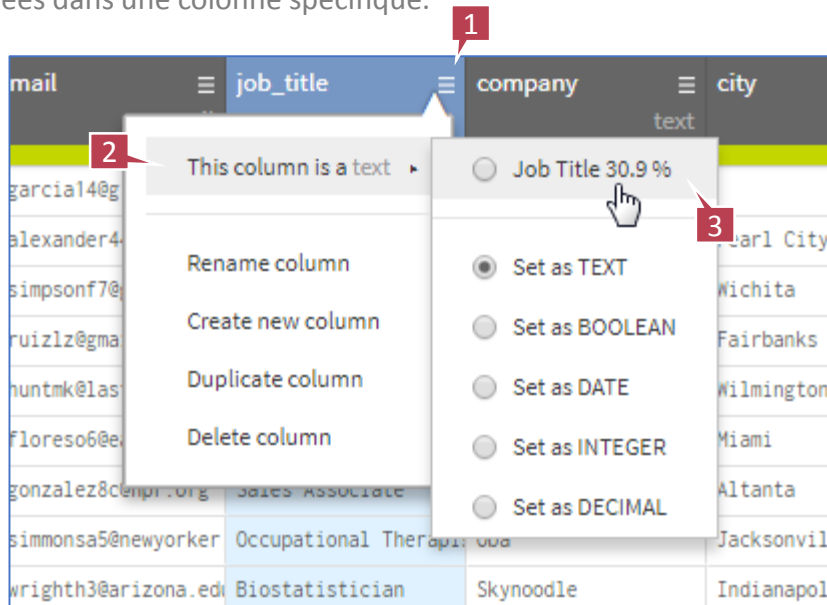
id	Name	last_name	email	job_title	company	city	state
	first name	last name	email			city	city
1	Kathryn	Garcia	kgarcia14@				
2	Jason	Alexander	jalexander44@gmail.c	Chemical Engineer	Abata	Pearl City	HI
3	Lillian	Simpson	lsimpson7@gmail.com	Desktop Support Tech	Camimbo	Wichita	KS
4	Walter	Ruiz	wruiz12@gmail.com	Geological Engineer	Yakitri	Fairbanks	AK
5	Joshua	Hunt	jhuntnk@last.fm	Financial Advisor	Oyope	Wilmington	DE
6	Mildred	Flores	mflores06@earthlink.	Nurse	Edgeblab	Miami	FL
7	Victor	Gonzalez	vgonzalez28c@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	Shuffletag	Racine	WI
15	Victor	Cox	vcocx9@virginia.edu	Librarian	Skalith	Bend	OR

- Vous pouvez créer et sauvegarder plusieurs préparations pour chaque jeu de données.
- Comme vous avez créé cette préparation à l'aide du bouton « Add Preparation », vous n'avez pas besoin de sauvegarder votre travail. Chaque nouvelle étape de préparation est automatiquement sauvegardée.
- N'oubliez pas que les données d'origine de votre jeu de données ne sont pas modifiées.

Type sémantique

Type sémantique :

Le type sémantique correspond à l'en-tête de la colonne ou à ce que représentent les données dans une colonne spécifique.



Le type sémantique suggéré pour la colonne Job Title est « text ». Il faut donc le convertir en un type qui ait plus de sens, « Job Title » dans notre cas.

1. Pour modifier le type sémantique, cliquez sur **la flèche en bas** dans l'en-tête de la colonne, puis choisissez le nouveau type sémantique.
2. Passez votre souris au dessus de **This column is a text**.
3. Choisissez **Job Title**.

La version « Enterprise Edition » de Talend Data Preparation vous permet de créer des types sémantiques personnalisés. Elle vous permet également de modifier ou supprimer les types sémantiques par défaut.

Talend Data Preparation suggère automatiquement les formats appropriés pour chaque colonne de vos jeux de données, ce qui vous aidera à mieux découvrir vos données. Toutefois, vous pouvez modifier ces suggestions à tout moment, suivant votre propre expérience.

Barre de qualité des données

En haut de chaque colonne, une barre mesure la qualité des données et indique par un code couleur le nombre de champs valides, vides ou invalides.

- Vert** – les données correspondent au format de la cellule
- Blanc** – cellules vides
- Orange** – les données ne correspondent pas au format de la cellule

id	Name	last_name	email	job_title
integer	first name	us county	email	text

Regardons de plus près la barre de qualité de la colonne adresse e-mail. En glissant le curseur sur chaque couleur, le nombre exact et le pourcentage des valeurs correspondant s'affichent.

- Vert** – 978 cellules ont un format valide
- Blanc** – 20 cellules sont vides
- Orange** – 2 cellules ont un format invalide

Pour sélectionner, supprimer ou vider les cellules ayant un format invalide, il suffit de cliquer sur la barre colorée. Cliquez sur la section orange, puis sélectionnez dans le menu déroulant « Select line with invalid values for E-Mail » afin de visualiser les adresses e-mail dont le format est invalide.

email	job_title
email	text
kgarcia14@g	
jaalexander44@gmail.com	Software Engineer

email	job_title
email	text
kgarcia14@g	
jaalexander44@gmail.com	Software Engineer

email	job_title
email	text
kgarcia14@g	
jaalexander44@gmail.com	Software Engineer

last_name	email	job_title
us county	email	

Select rows with invalid values for email

Clear the cells with invalid values

Delete the rows with invalid cell

N'oubliez pas de supprimer le filtre pour revenir à la liste complète.

Filters

Add a filter ...

email: rows with invalid values

Manipulation de texte basique

Filtrer et corriger les données

Pour filtrer les lignes invalides :

1. Cliquez sur l'en-tête de la colonne STATE

2. En bas à droite, vous pouvez sélectionner et visualiser un graphique des modèles de données (Pattern). En glissant le curseur sur chaque ligne, des analyses quantitatives vont apparaître. La ligne du haut indique que 991 enregistrements contiennent un code à deux lettres (correspondant à un État). **Si vous cliquez sur une des barres, vous visualiserez uniquement les enregistrements correspondants. Pour supprimer le filtre, cliquez sur « x » dans l'onglet qui s'ouvre au-dessus de la colonne.**

3. Cliquez sur la section orange de la barre de qualité.

4. Cliquez sur « Selects rows with invalid values for STATE ».

5. 7 lignes contenant des informations invalides s'affichent.

The screenshot illustrates the process of filtering and correcting data in Talend Data Preparation. It shows a data table with columns 'city', 'state', and 'date'. The 'state' column is highlighted, and a dropdown menu is open, showing options to 'Select rows with invalid values for state', 'Clear the cells with invalid values', and 'Delete the rows with invalid cell'. A pattern analysis chart is also visible, showing the distribution of values in the 'state' column. The chart has a 'PATTERN' tab selected, and a bar chart shows the frequency of different state codes. A red box highlights the 'PATTERN' tab, and a red box highlights the 'Select rows with invalid values for state' option in the dropdown menu.

city	state	date
Pearl City	HI	22/11/2015
Wichita	KS	2/28/2015
Fairbanks	AK	7/15/2015
Wilmington	DE	3/16/2015
Miami	FL	10/15/2015
Atlanta	GA	17-12-2014
Jacksonville	FL	17-12-2015
Indianapolis	IN	01/01/2016 10:00:0

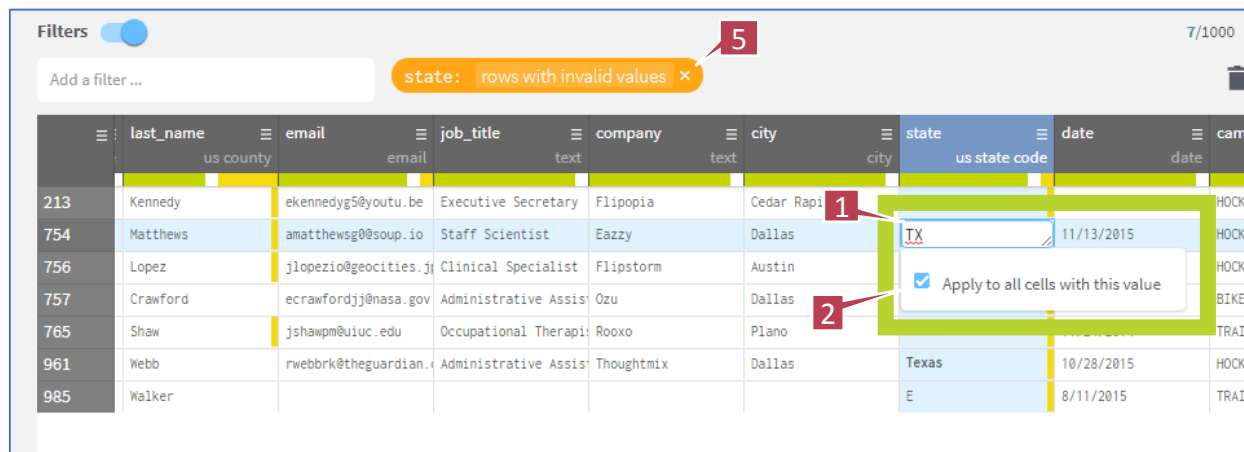
Ici nous avons nettoyé et modifié la valeur des champs ayant un format invalide.

Vous apprendrez comment utiliser les graphiques pour filtrer les données mais aussi pour modifier des valeurs directement dans la grille des données.

Manipulation de texte basique

Filtrer et corriger les données

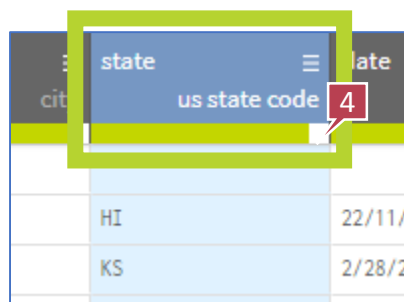
1. Pour modifier le texte d'un champ, **double-cliquez sur une des cellules** contenant « Texas ». Changez Texas en TX (son code). **N'appuyez PAS sur la touche Entrée pour le moment !**
2. Au-dessous de la cellule que vous modifiez, cochez « **Apply to all Cells with this value** » pour appliquer la modification à toutes les cellules de la même valeur. **MAINTENANT vous pouvez appuyer sur Entrée !** Vous venez de changer toutes les cellules avec la valeur « Texas » en « TX ».
3. Il nous reste deux lignes de données invalides. Regardez la liste des fonctions disponibles et **choisissez celle que vous souhaitez** utiliser pour corriger le code invalide.
4. Une fois toutes ces actions effectuées, la **barre de qualité des données** au-dessous de STATE sera uniquement **verte et blanche**.
5. Cliquez sur « x » dans la boîte **recherche/filtre** pour annuler le filtre actuel et revenir à la liste de données complète.



Filters ☒ 7/1000

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

	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	HOCK
754	Matthews	amathewsg0@soup.io	Staff Scientist	Eazzy	Dallas	TX		HOCK
756	Lopez	jlopezio@geocities.jp	Clinical Specialist	Flipstorm	Austin			HOCK
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	HOCK
985	Walker					E	8/11/2015	TRAI



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

Recettes

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

Chaque fonction utilisée a été ajoutée à la recette. La dernière étape nous dit que tous les champs ayant Texas comme État ont été changés en TX.

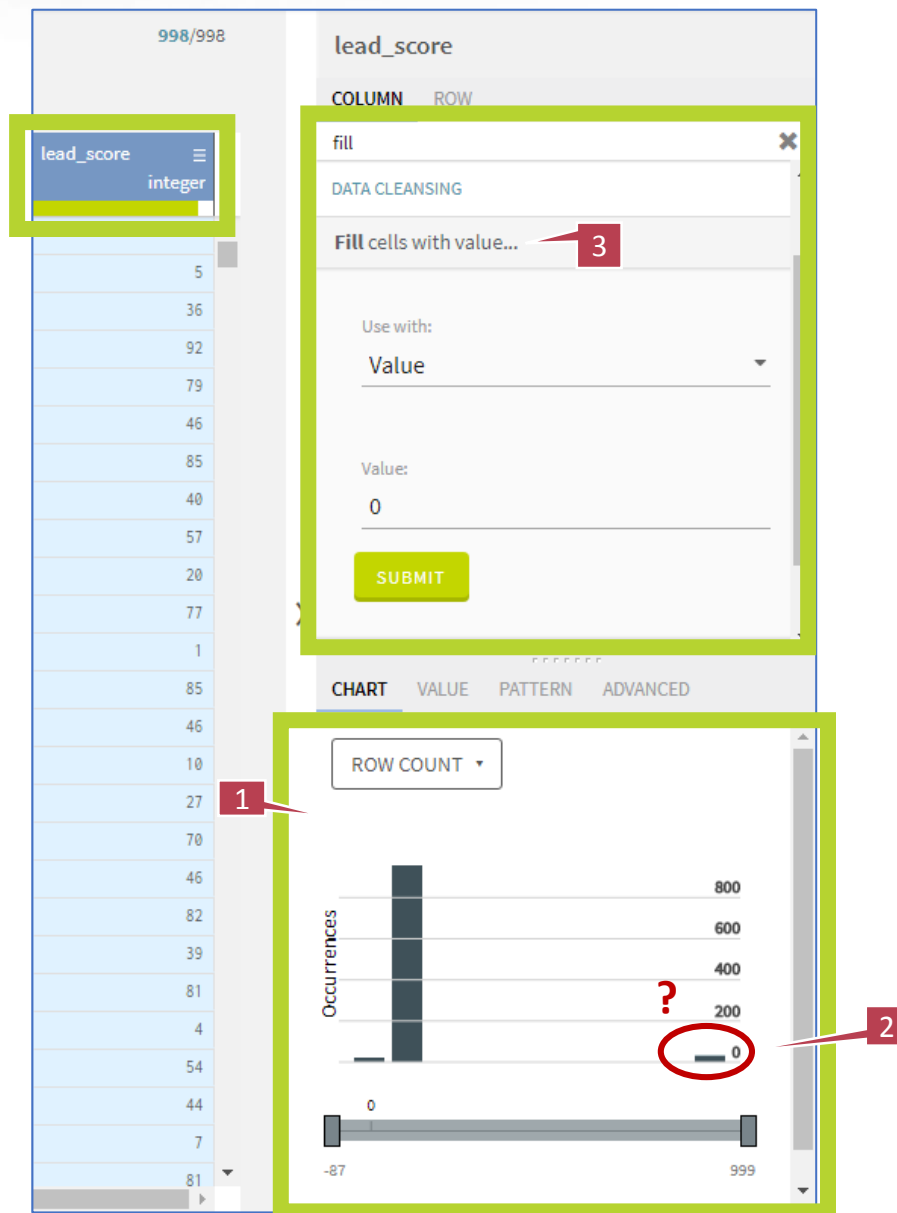
Manipulation numérique basique

Filtrer et corriger les données

Maintenant, passons à la colonne **LEAD_SCORE**.

1. Il s'agit ici de champs de nombres entiers, mais **l'histogramme** à droite de l'écran nous dit que les données sont faussées par des valeurs plus grandes.
2. Cliquez sur la **barre bleue** tout à droite dans le graphique : 31 cellules ont la valeur 999. Il semblerait que la valeur par défaut soit réglée à 999. Cette fois, nous utiliserons la fonction « **Fill Cell with Value ...** » pour modifier la valeur.
3. Tapez « **Fill** » dans la boîte de recherche en haut à droite de l'écran, puis sélectionnez « **Fill Cell with Value...** ». Réglez la valeur à 0 et cliquez sur **Submit**.

Ici nous avons nettoyé et modifié des valeurs aberrantes dans un champ numérique. Vous apprendrez comment utiliser les graphiques pour filtrer les données, mais aussi pour modifier des valeurs directement dans la grille des données.



Manipulation numérique basique

Filtrer et corriger les données

Champ LEAD_SCORE (suite)

1. En observant de plus près le graphique se rapportant à la colonne LEAD_SCORE, vous remarquerez des valeurs négatives.
2. Puisqu'il est impossible d'avoir des scores négatifs, il faut les supprimer. Dans le menu des fonctions suggérées, sélectionnez « Calculate Absolute Value ». Cette fonction permet de garder la valeur des scores tout en éliminant le signe négatif.

The screenshot displays the Talend Data Preparation 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, the 'lead_score' column configuration panel is visible. It shows a list of suggestions for functions to apply to the column. The 'Calculate absolute value' option is highlighted with a red box and a red arrow pointing to it, labeled with a red '2'.

Below the suggestions list, a 'CHART' tab is active, showing a bar chart of 'occurrences' for the 'lead_score' column. The chart shows a distribution of values, with a red box and a red arrow pointing to the 'Calculate absolute value' option in the suggestions list, labeled with a red '1'.

The chart also shows a range from 'Min -87' to 'Max 0'.

Nettoyage et formatage de dates

Filtrer et corriger les données

Occupons-nous maintenant du champ Date.

1. Cliquez sur **le champ Date, puis sur Pattern à droite de l'écran**. Vous pourrez visualiser aisément tous les différents formats de date et de masquage utilisés. Certaines dates sont au format européen, d'autres au format américain, certaines contiennent des slashes, d'autres des tirets.
2. Pour standardiser les dates, cliquez sur **« Change Date Format » dans la liste des fonctions suggérées**. Sélectionnez un format parmi ceux proposés ou insérez celui de votre choix, puis cliquez sur **Submit**.

Combien de fois sommes-nous confrontés à des tableaux comportant toute sorte de formats et standards de dates extravagants ? Nous savons tous qu'Excel peut reformater les champs date, mais quand ils présentent un mélange de formats et différents masquages, Excel ne gère plus !

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. On the right, the 'date' column is selected, and the 'SUGGESTIONS' list is open. The 'Change date format...' option is highlighted with a green box and a red arrow pointing 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 'custom'. The 'Your format' field contains 'MM.dd.yyyy'. A red box with the number '3' is next 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

Nettoyage et formatage de dates

Modifier la recette, c'est facile.

1. Dans la recette, à gauche, **surlignez la dernière action utilisée.**
2. Cliquez sur la flèche, puis dans le menu déroulant qui s'affiche choisissez **other (design your custom pattern)**. Tapez **dd-
MMMM-yyyy** pour modifier le **format** (attention aux minuscules et aux majuscules, cette fonction y est très sensible).
3. Cliquez sur **Submit**, les **modifications seront alors appliquées.** Vous pouvez supprimer une formule de la liste (cliquez sur la corbeille) ou l'enregistrer (cliquez sur la coche verte).
4. Vous pouvez **réordonner les étapes de votre recette en les glissant.** Cela vous fera gagner du temps si, par exemple, vous réalisez qu'une colonne sur laquelle vous avez appliqué une formule contenait encore des données invalides restant à nettoyer.

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	bwright3@arizona.edu	Biostatistician	Skynoodle
10	frodriqueznc@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

Your format: dd-MMMM-yyyy

SUBMIT

Le masquage de données

Vous pouvez facilement masquer des données sensibles :

1. Cliquez sur la colonne **EMAIL** pour sélectionner son contenu.
2. Dans la liste des fonctions, recherchez **Mask data (Obfuscation)**.
3. Cliquez dessus pour appliquer la fonction sur la colonne **EMAIL**.
4. Tous les caractères avant @ sont remplacés par XXX, alors que le reste ne change pas. C'est l'effet de la fonction de masquage de données sur les cellules dont le type sémantique est l'email. Mais l'effet du masquage de données sera différent selon le type sémantique de la colonne.

The screenshot shows the Talend Data Preparation interface with a table titled 'Customer Marketing Leads Preparation'. The table has columns: id, Name, last_name, email, title, company, city, state, date, campaign_id, and lead_score. The 'email' column is highlighted with a green box and labeled '1'. A context menu is open over the 'email' column, showing the 'Mask data (Obfuscation)' function selected, labeled '3'. The function's configuration window is also visible, showing the 'email' column selected and the 'Mask data (Obfuscation)' function applied. The result of the function is shown in the table, where the email addresses are masked with 'XXXXXXX' followed by the domain part, labeled '4'.

Lorsque vous manipulez des données sensibles telles que des noms, des adresses, des numéros de carte de crédit ou de sécurité sociale, vous pouvez avoir besoin de masquer ces données. Afin de protéger les données d'origine, vous pouvez utiliser la fonction de masquage de données afin de générer des alternatives fonctionnelles.

Le Data Blending

Data Blending

Le Data Blending consiste à combiner les données issues de différentes sources. Cette fonction vous permet d'importer des données d'un jeu de données et de les ajouter à celui sur lequel vous êtes en train de travailler.

1. Cliquez sur l'icône Data Blending.
2. La liste comprenant les jeux de données que vous avez enregistrés et d'autres fichiers pré-chargés est disponible en cliquant sur l'icône +.
3. Cochez « Business Unit Regions with States » puis cliquez Add.

1

2

3

3b

Le Data Blending

Data Blending (suite)

1. Cliquez sur la **colonne que vous souhaitez importer et combiner**, dans ce cas la **colonne State**, dans le jeu de données en cours d'utilisation.
2. Ajoutez les régions en cliquant sur **Add to Dataset** au-dessous de l'en-tête de la colonne « Region ».
3. **Positionnez-vous sur Confirm** afin de prévisualiser les modifications qui seront affichées en jaune. Pour appliquer ces modifications, **cliquez sur Confirm**.

The screenshot shows the Talend Data Preparation interface. 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. The 'state' column is highlighted in blue. A red callout '1' points to the 'state' column header. A red callout '2' points to the 'Add to Dataset' button in the 'Region' column header. A red callout '3b' points to the 'Confirm' button in the bottom right corner. The interface also includes a 'Filters' section, a 'SUGGESTIONS' panel, and a 'CHART' panel on the right.

	first_name	last_name	email	job_title	company	city	state	us_state_code	Region	date	campaign_id
2	Alexander	xxxxxx@gmail.com	Chemical Engineer	Abata	Pearl City	HI	West	11.22.2015	HOOKEY_Y15Q01_cant		
3	Simpson	xxxxxx@gmail.com	Desktop Support Tech	Canibo	Wichita	KS	Mid West	02.28.2015	RUN_Y14Q02_deal		
4	Ruiz	xxxxxx@gmail.com	Geological Engineer	Yakitri	Fairbanks	AK	West	07.15.2015	TRAIL_Y14Q04_purr		
5	Hunt	xxxxxx@last.fm	Financial Advisor	Oyope	Wilmington	DE	North East	03.16.2015	HOOKEY_Y14Q02_mode		
6	Flores	xxxxxx@earthlink.net	Nurse	Egeblab	Miami	FL	South East	10.15.2015	HOOKEY_Y15Q04_chur		
7	Gonzalez	xxxxxx@npr.org	Sales Associate	Ntag	Atlanta	GA	South East	12.17.2014	TRAIL_Y15Q03_hold		
8	Simmons	xxxxxx@newyorker.com	Occupational Therapist	Oba	Jacksonville	FL	South East	12.17.2015	TRAIL_Y14Q03_moon		
9	Wright	xxxxxx@arizona.edu	Biostatistician	Skynoodle	Indianapolis	IN	Mid West	01.01.2016	TRAIL_Y15Q04_rossy		
10	Rodriguez	xxxxxx@fotki.com	Director of Sales	Eidel	Anchorage	AK	West	07.06.2015	BTXK_Y14Q02_hurt		
11	Peterson	xxxxxx@sohu.com	Research Nurse	Gabcube	Las Vegas	NV	West	03.16.2015	HOOKEY_Y15Q02_boos		
12	Martin	xxxxxx@java.com	Speech Pathologist	Zoomcast	Nampa	ID	West	12.09.2014	SKI_Y15Q02_vied		
13	Sullivan	xxxxxx@brycos.com	Automation Specialist	Bluezoom	Bridgeport	CT	North East	06.01.2015	SKI_Y14Q03_yack		
14	Gonzales	xxxxxx@apple.com	Automation Specialist	Shuffletag	Racine	WI	Mid West	11.02.2015	HOOKEY_Y14Q04_roan		
15	Cox	xxxxxx@virginia.edu	Librarian	Skalth	Bend	OR	West	06.04.2015	TRAIL_Y15Q04_hays		
16	Wilson	xxxxxx@va.gov	Actuary	Rhyloo	Manhattan	NY	North East	11.02.2015	TRAIL_Y14Q04_fete		
17	Arnold	xxxxxx@youtube.com	Senior Editor	Tazzy	Columbus	GA	South East	12.25.2014	HOOKEY_Y15Q01_file		

State | **Region**

us state code | city

1 | WA | West

2 | MT | West

3 | OR | West

4 | ID | West

5 | WY | West

6 | CA | West

7 | NV | West

8 | UT | West

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.

CONFIRM

Regrouper et standardiser

Regrouper les données

Les fonctions pour regrouper et standardiser les données vous permettent de localiser des cellules ayant un contenu texte similaire afin de les réunir et d'harmoniser le texte.

1. Cliquez sur l'en-tête de la colonne **JOB_TITLE**.
2. Le graphique en bas à droite montre la grande quantité de dénominations de poste similaires dans le fichier. Afin d'harmoniser, nous devons tout d'abord regrouper les dénominations similaires.
3. Dans la boîte de recherche, tapez « group ».
4. Cliquez sur « Find and Group Similar Text » pour trouver et regrouper les valeurs texte similaires.

The screenshot displays the Talend Data Preparation interface. On the left, a data table with columns: job_title, company, city, state, and us state code. The 'job_title' column is selected, indicated by a red '1'. On the right, the 'Find and Group Similar Text' function is applied to the 'job_title' column, indicated by a red '3'. The search bar contains the text 'group', indicated by a red '4'. Below the search bar, a bar chart shows the frequency of job titles, with 'Occupational Therapist' being the most frequent, indicated by a red '2'.

job_title	company	city	state	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	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	

job_title

COLUM 3 OW

group

SPLIT

Extract string parts...

STRINGS ADVANCED

Find and group similar text...

CHART VALUE PATTERN ADVANCED

ROW COUNT

0 5 10 15 20

(EMPTY)

Occupational Therapist

Database Administrator

VP Marketing

Financial Advisor

Web Designer

Software Consultant

Human Resources Manager

Librarian

Senior Financial Analyst

Geological Engineer

Structural Engineer

VP Sales

VP Product Management

Environmental Tech

Trouver et regrouper des données texte similaires

Regrouper les données

Regrouper et standardiser (suite)

1. **Toutes les dénominations de poste similaires sont regroupées** dans la deuxième colonne.
2. La troisième colonne suggère la dénomination de poste qui pourrait **remplacer** celles de la deuxième colonne. Vous pouvez utiliser **le menu déroulant pour choisir une autre dénomination ou bien insérer celle de votre choix**.
3. Si vous ne souhaitez pas modifier une dénomination spécifique, **décochez** la case devant la dénomination en question.
4. Si vous ne souhaitez pas modifier un groupe spécifique de dénominations de poste, **décochez** la case dans la première colonne.
5. Cliquez sur **Submit** quand vous avez terminé.

FIND AND GROUP SIMILAR TEXT

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

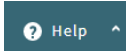
<input checked="" type="checkbox"/>	These values have been found	1	This value will be kept
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Health Coach <input checked="" type="checkbox"/> Health Coach1	3	Replace value: <input type="text" value="Health Coach"/>
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Administrative Assistant <input checked="" type="checkbox"/> Administrative Officer		Replace value: <input type="text" 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: <input type="text" value="Accountant"/>

SUBMIT

5

Commentaires

Vos commentaires sont très importants pour nous. Nous souhaitons savoir si nos services répondent à vos besoins et si nous vous les proposons de manière efficace.

Pour laisser un commentaire, cliquez sur la flèche à côté de  et, dans le menu déroulant, sélectionnez **Feedback**. Complétez le formulaire avec votre adresse e-mail et vos observations. N'hésitez pas à utiliser également les liens vers le forum et la base de connaissances présents dans le formulaire.

×

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



Maintenant, vos données sont prêtes :

- **pour l'analyse** : vos données sont nettoyées et standardisées, et vous avez exporté les résultats de votre préparation. Vous pouvez par exemple analyser le potentiel de vos prospects par date ou par région dans Excel ou dans Tableau.
- **pour l'intégration** : les données sont nettoyées et formatées, vous pouvez les charger dans une application de CRM ou d'automatisation du marketing, telle que Marketo ou Salesforce.com.

La bonne nouvelle c'est que...

Avec Talend, les données sont à un clic de vos tâches quotidiennes.

Et après ?

Maintenant que vous avez appris à ajouter vos jeux de données dans l'application, à effectuer vos propres préparations... vous pouvez transformer vos tâches quotidiennes en activités guidées par les données.

