

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



Cloud Dataprep Demo: Inconsistent company names

Standardising values in a text column

Celia Muriel – celiamuriel.com
[Post](#), [GitHub repository](#).

Introduction	1
Prepare demo environment	2
Source file in Cloud Storage	2
Target BigQuery table	2
Enable and setup Cloud Dataprep	2
Remove inconsistent values	2
Standardise company names	2
Recipe and flow	47
Other option	47
References	50
Trifacta	50
Other	50

Introduction

This demo shows how to remove inconsistent data in a [CSV file](#) and load it to BigQuery. The source file has company names in a column. Their values are inconsistent (VANILLA Ltd, *** VANILLA LTD ***, vanilla ltd, vanila ltd, Vanilla Ltd., etc.). We are going to standardise the company names before uploading them to the target table.

There is no one single technique to achieve this task. We are going to use several out-of-the-box solutions implemented in Cloud Dataprep (by Trifacta), including clustering of values based on fuzzy matching.

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



This demo was done on June 13th, 2020, with the Generally Available features on the different services used for this exercise.

Prepare demo environment

Source file in Cloud Storage

Create a Cloud Storage bucket for the demo. Upload to the bucket the [source file](#).

Target BigQuery table

Create the BigQuery dataset and table as shown below.

```
bq --location EU mk \  
--dataset \  
[PROJECT_ID]:vanilla  
  
bq mk \  
--table \  
[PROJECT_ID]:vanilla.companies \  
company_name:STRING,company_id1:STRING,company_id2:STRING,company_id3:STR  
ING,company_id4:STRING,country:STRING,town:STRING,zipcode:STRING
```

Enable and setup Cloud Dataprep

If we want to use Cloud Dataprep, we need our account to grant permissions to any of these roles: project editor, project owner or dataprep.user.

Then we need to [enable Cloud Dataprep](#) and set it up. If you are running your ecosystem outside the US, and you have requirements to be in a certain region, make sure that the Cloud Storage buckets Cloud Dataprep are set up according to your requirements.

Remove inconsistent values

Standardise company names

Once we have enabled and set up Cloud Dataprep, we create a flow.

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



Flows Create... ...

All Flows Owned by me Shared with me

Search...

Create a flow to wrangle your data.

Create Flow

Flow Name: Vanilla Flow

Flow Description: Standardise company names

Cancel Create

Then we add the companies.csv file as a dataset.

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



Vanilla Flow
Standardise company names

Import data before wrangling in this Flow.

Import & Add Datasets

Import Data and Add to Flow

Choose a file or folder

GCS

Create Dataset with Parameters

0 New Datasets

Choose data to import.

NAME SIZE LAST UPDATED

dataprep-staging-fb...
dataprep-staging-fb...
vanilla_bucket

Import & Add to Flow Cancel

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



Import Data and Add to Flow

Choose a file or folder
GCS / vanilla_bucket [Create Dataset with Parameters](#)

0 New Datasets

Choose data to import.

NAME SIZE LAST UPDATED

companies.csv 2.79MB Today at 8:28 PM

Import & Add to Flow **Cancel**

Import Data and Add to Flow

Choose a file or folder
GCS / vanilla_bucket [Create Dataset with Parameters](#)

1 New Dataset [Clear All](#)

companies.csv

Add a Description

RBC Country	RBC	Co
ES	INGETEAM POWER TE	

[Edit settings](#)

Import & Add to Flow **Cancel**

We must create a recipe to clean and prepare the data in companies.csv to upload it to BigQuery.

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



Screenshot of the Vanilla Flow interface showing a dataset named "companies.csv". The dataset contains a single column of company names, all of which are identical: "INGETEAM POWER TECHNOLOGY".

The "Details" panel shows the following information:

RBC Country	RBC	Company
ES	INGETEAM POWER TECHNOLOGY	INGETEAM POWER TECHNOLOGY
ES	INGETEAM POWER TECHNOLOGY	INGETEAM POWER TECHNOLOGY
ES	INGETEAM POWER TECHNOLOGY	INGETEAM POWER TECHNOLOGY
ES	INGETEAM POWER TECHNOLOGY	INGETEAM POWER TECHNOLOGY
ES	INGETEAM POWER TECHNOLOGY	INGETEAM POWER TECHNOLOGY
ES	INGETEAM POWER TECHNOLOGY	INGETEAM POWER TECHNOLOGY

Type: GCS
Location: gs://vanilla_bucket/companies.csv
File Size: 2.79MB
Size: 8 columns · 2 types
Updated: Today at 8:37 PM
Created: Today at 8:37 PM

A red arrow points to the "Add" button in the "Dataset details" section. Another red arrow points to the "Recipe" dropdown menu, which includes options like "Join" and "Union".

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



The screenshot shows the Vanilla Flow interface. On the left is a sidebar with various icons. The main area displays a flow diagram where a dataset named 'companies.csv' is connected to a recipe named 'companies'. The recipe details panel on the right shows a 'Edit Recipe' button, which is highlighted with a red arrow. Other buttons include 'Add' and three dots. The panel also includes tabs for 'Recipe' and 'Data', a 'Steps Preview' section with a note about no steps, and metadata like 'Steps: 0', 'Updated: Today at 8:40 PM', and 'Created: Today at 8:40 PM'.

When we edit the recipe for the first time, it shows us the data in the CSV file. Dataprep allows to quickly analyse the data and add the transformations easily.

The screenshot shows the Google Cloud Dataprep interface. The top navigation bar includes 'VANILLA FLOW >', 'companies', 'Full Data', a search bar, and a 'Run Job' button. The main area is a data preview table with the following columns: RBC, Country, Company Name, Company ID 1, Company ID 2, and Company ID 3. The preview shows several rows of data with histograms above each column indicating the distribution of values. The bottom status bar indicates '8 Columns', '49,788 Rows', and '2 Data Types'.



In this demo, we are going to clean and prepare three fields, but we are going to focus on the company name (**Vanilla** in companies.csv). See that when we open the CSV file before standardising the companies, there are 49,788 rows in the file and 45,032 different company names.

VANILLA FLOW > companies Full Data

Run Job

Company Name

165 Categories 45,032 Categories 22,191 Categories 28 Categories 33 Categories

RBC	Country	RBC	Company Name	RBC	Company ID 1	RBC	Company ID 2	RBC	Company ID 3
ES	INGETEAM POWER TECHNOLOGY SA SERVICE	A95663852							
ES	INGETEAM POWER TECHNOLOGY SA PANELS	A95663852							
ES	INGETEAM POWER TECHNOLOGY SA ENERGY	A95663852							
ES	INGETEAM POWER TECHNOLOGY SA INDUSTRY	A95663852							
ES	INGETEAM POWER TECHNOLOGY SA MARINE	A95663852							
ES	INGETEAM POWER TECHNOLOGY SA ELECTRONICS	A95663852							
ES	INGETEAM POWER TECHNOLOGY SA TECHNOLOGY	A95663852							
ES	INGETEAM POWER TECHNOLOGY SA POWER PLANTS	A95663852							
CL	Ingeteam SpA	76198448-9							
AU	Ingeteam Australia Pty Ltd	51166870168							
PA	Ingeteam Panamá SA	2619428-1-836502							
RO	Ingeteam Service S.R.L.								
ES	QUANTUM RENEWABLE ENERGY SL	B71219323							
MX	"Tai-Durango Dos, S. A. P. I. de C. V."	TDD1204178Y8							
MX	"Tai-Durango Cuatro, S.A.P.I. de C.V."	TDC120417R14							
PH	Ingeteam Philippines Inc.	8996479000							
ES	BIZKAIA BUSINESS CAPITAL SL	B95810552							
ES	BIZKAIA BUSINESS CAPITAL 1-SL EN CONSTITUCIÓN	B95813069							
ES	BIZKAIA BUSINESS CAPITAL 2-SL EN CONSTITUCIÓN	B95826236							
UY	Ingeteam Uruguay S.A.	21-762707-0016							
ES	GLOBAL INNOVATION SLU	B97051600							

8 Columns 49,788 Rows 2 Data Types

If we scroll right, we see that the quality bar for the ZIP code shows mismatched values. This is due to the fact that most of the ZIP codes in this file are Spanish, which means they are made of 5 digits. However, some companies are outside Spain and the ZIP codes may be alphanumeric. So we change the data type to string.

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

Details X

Company ID 3	RBC	Company ID 4	RBC	Country1	RBC	ZIP Code	RBC	ZIP Code
es	20,050 Categories	16,469 Categories		Albacete				
ESA95663852		Seasma		Sarriguren				
ESA95663852		Zamudio		Zamudio				
ESA95663852		Zamudio		Zamudio				
ESA95663852		Zamudio		Zamudio				
ESA95663852		Zamudio		Miñano Mayor				
ESA95663852		Zamudio		Las Condes				
ESA95663852		Zamudio		North Wollongong				
ESA95663852		Zamudio		Distrito de Panamá				
RO34091550		Zamudio		"Bucuresti, Sector 2"				
		Zamudio		Orkien				
		Zamudio		CIUDAD DE MÉXICO				
		Zamudio		CIUDAD DE MÉXICO				
		Zamudio		Makati City				
		Zamudio		Zamudio				
		Zamudio		Zamudio				
		Zamudio		Montevideo				

8 Columns 49,788 Rows 2 Data Types

More options: Rename, Change type, Move, Hide, Format, Calculate, Create column from examples, Group by, Pivot, Restructure, Filter rows, Replace, Standardize, Extract, Split column, Column Details, Show related Steps in Recipe.

VANILLA FLOW > companies > Full Data

Details X

D 1	RBC	Company ID 2	RBC	Company ID 3	RBC	Company ID 4	RBC	Country1	RBC	ZIP Code
		28 Categories		33 Categories		20,050 Categories		16,469 Categories		12,026 Categories
		ESA95663852		ESA95663852		ESA95663852		Albacete		2006
		ESA95663852		ESA95663852		ESA95663852		Seasma		31293
		ESA95663852		ESA95663852		ESA95663852		Sarriguren		31621
		ESA95663852		ESA95663852		ESA95663852		Zamudio		48170
		ESA95663852		ESA95663852		ESA95663852		Zamudio		48170
		ESA95663852		ESA95663852		ESA95663852		Zamudio		48170
		ESA95663852		ESA95663852		ESA95663852		Miñano Mayor		1510
		ESA95663852		ESA95663852		ESA95663852		Las Condes		7550000
		ESA95663852		ESA95663852		ESA95663852		North Wollongong		2500
		ESA95663852		ESA95663852		ESA95663852		Distrito de Panamá		"Bucuresti, Sector 2"
		ESA95663852		ESA95663852		ESA95663852		Orkien		20335
		ESA95663852		ESA95663852		ESA95663852		CIUDAD DE MÉXICO		31160
		ESA95663852		ESA95663852		ESA95663852		CIUDAD DE MÉXICO		11550
		ESA95663852		ESA95663852		ESA95663852		Makati City		11550
		ESA95663852		ESA95663852		ESA95663852		Zamudio		1200
		ESA95663852		ESA95663852		ESA95663852		Zamudio		48170
		ESA95663852		ESA95663852		ESA95663852		Zamudio		48170
		ESA95663852		ESA95663852		ESA95663852		Montevideo		11200

8 Columns 49,788 Rows 1 Data Type

In the source application when someone wanted to indicate that a company name shouldn't be used, they have written "do not use" in Spanish and Italian within the company name. We are

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



going to remove all those strings as the company names are valid in our historical information and we should use an end date to mark the validity of a company. First we look for those rows.

VANILLA FLOW >
Companies < Full Data

Run Job

Columns: RBC Country Company Name RBC Company ID 1 RBC Company ID 2 RBC Company ID 3

Rows: 165 Categories, 45,032 Categories, 22,191 Categories, 28 Categories, 33 Categories

Company Name Data:

Country	Company Name	RBC	Company ID 1	RBC	Company ID 2	RBC	Company ID 3
ES	INGETEAM POWER TECHNOLOGY SA SERVICE	A95663852					
ES	INGETEAM POWER TECHNOLOGY SA PANELES	A95663852					
ES	INGETEAM POWER TECHNOLOGY SA ENERGY	A95663852					
ES	INGETEAM POWER TECHNOLOGY SA INDUSTRY	A95663852					
ES	INGETEAM POWER TECHNOLOGY SA MARINE	A95663852					
ES	INGETEAM POWER TECHNOLOGY SA ELECTRONICS	A95663852					
ES	INGETEAM POWER TECHNOLOGY SA TECHNOLOGY	A95663852					
ES	INGETEAM POWER TECHNOLOGY SA POWER PLANTS	A95663852					
CL	Ingeteam SpA	76198448-9					
AU	Ingeteam Australia Pty Ltd	51166870168					
PA	Ingeteam Panamá SA	2619428-1-836502					
RO	Ingeteam Service S.R.L.						
ES	QUANTUM RENEWABLE ENERGY SL	B71219323					
MX	"Tai Durango Dos, S.A. P.I. de C."	TDD1204178Y8					
MX	"Tai Durango Cuatro, S.A.P.I. de C.V."	TDC120417R14					
PH	Ingeteam Philippines Inc.	8996479000					
ES	BIZKAIA BUSINESS CAPITAL SL	B95810552					
ES	BIZKAIA BUSINESS CAPITAL 1 SL EN CONSTITUCIÓN	B95813669					
ES	BIZKAIA BUSINESS CAPITAL 2 SL EN CONSTITUCIÓN	B95826236					
UY	Ingeteam Uruguay S.A.	21-762707-0016					
ES	GLOBAL INSTITUTE SLU	B97051602					

8 Columns 49,788 Rows 1 Data Type

We search for “no usar” (do not use).

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

Columns Rows

Rows

RBC	Country	RBC	Company Name	RBC	Comp
					no user
165 Categories	45,032 Categories				
ES	"NO-USAR - ACERINOX, S.A."				
MX	(NO-USAR) OCEAN EUROPE SA DE CV				
PA	Solarcentury (NO-USAR)				
ES	ETECNIC SCP (NO-USAR)				
RU	no user				
MA	NO-USAR-ACWA POWER BOUDOUR				
ES	Montajes y Mantenimientos 2020 SL NO-USAR				
ES	NO-USAR SISTECA INNOVATION SL				
ES	"CROVI, S.A. (NO-USAR)"				
ES	Noratel-Spain S.L. (NO-USAR)				
ES	"Urfiltr, S.L. NO-USAR NO-USAR"				
ES	"NO-USAR(BZZ MOBILIARIO,S.L.)"				
ES	"Releco, S.A. ***NO-USAR***VER DISAILECO 206949"				
GB	Southco-EURpe Ltd.(NO-USAR)				
ES	NO-USAR * EPCOS ELECTRONIC COMP. SUSTITUIR POR 208102				
IT	Falco Electronics (No user)				
ES	"NO-USAR - PROSEGUR CIA. SEGURIDAD, S."				
NL	NO-USAR CADENCE SOLUTIONS PROVIDER				
GB	"Southco Manufacturing, Ltd.(NO-USAR)"				
US	NO-USAR * KEMET Electronics Corp				
DE	NO-USAR CO-MODULS AG				

8 Columns 50 Rows 1 Data Type

We highlight one of the “no user” occurrences. A Suggestions wizard shows on the right. We choose Replace with “”. A yellow column appears which shows how the company names will look after the replacement. We click on “add”.

VANILLA FLOW > companies > Full Data

Source to be dropped Preview

Suggestions

Count values matching See all

- 'NO-USAR'
- '(upper)+ (upper)+'
- 'NO-USAR' starting after `` ending before `'

Split on values matching See all

- 'NO-USAR'
- '(upper)+ (upper)+'
- 'NO-USAR' starting after `` ending before `'

Replace

'NO-USAR' with " in Company Name

Add

{upper}+ {upper}+ with " in Company Name

Extract list of values See all

- matching '(alpha)+*(alpha)+'
- matching '({upper}+)*{upper}+*

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

New Step Recipe X

1 Change ZIP Code type to String
2 Replace matches of 'NO USAR' from Company Name with "

RBC	Country	RBC	Company Name	RBC	Company ID 1	RBC	Company ID
							(1)
	165 Categories		45,032 Categories		22,191 Categories		28 Categories
RU			no user				
IT			Falco Electronics (No user)		108209990		
ES			"no user") INGETEAM Hydro, S.L."				
ES			No user CORPORACIÓN EÓLICA DE ZARA				

8 Columns 4 Rows 1 Data Type

We click on the filter again and on “Clear all filters”.

VANILLA FLOW > companies > Full Data

New Step Recipe X

1 Change ZIP Code type to String
2 Replace matches of 'NO USAR' from Company Name with "

Rows

no user

Clear all filters

8 Columns 4 Rows 1 Data Type



We continue searching for the “do not use” tags within the company name, and we remove them as done with “no usar”. Note that Cloud Dataprep is sensitive to capitalization. If we search for “do not use”, we find the records which have “DO NOT USE” and “do not use”. We must highlight and remove the occurrence in capital letters and the one in lowercase as well.

The screenshot shows the Google Cloud Dataprep interface. On the left, there's a sidebar with various icons. The main area displays a data preview for a flow named 'VANILLA FLOW > companies Full Data'. The preview shows two columns: 'RBC' and 'Country' on the left, and 'RBC' and 'Company Name' on the right. A filter dialog is open over the preview, with the 'Rows' tab selected. The search bar contains the text 'no usar'. Below the search bar, there are buttons for 'Clear all filters' and a gear icon. To the right of the preview, the 'Recipe' section is visible, showing two steps:

- 1 Change ZIP Code type to String
- 2 Replace matches of `NO USAR` from Company Name with "

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies >
[Full Data](#)

Recipe

Columns

165 Categories 45,031 Categories

RBC	Country	RBC	Company Name
			No usar
			Falco Electronics (No usar)
			No usar: CORPORACIÓN EÓLICA DE ZARA

8 Columns 2 Rows 1 Data Type

1 Change ZIP Code type to String
 2 Replace matches of 'NO USAR' from Company Name with ''
 3 Replace matches of 'no usar' from Company Name with ''

VANILLA FLOW > companies >
[Full Data](#)

Recipe

Columns

165 Categories 45,031 Categories

RBC	Country	RBC	Company Name
			no utilizad
			"NO UTILIZAR -- MICROBERRI
			"NO UTILIZAR -- Microsoft Ireland
			ESN0071290A
			"NO UTILIZAR -- Bergen Engines AS

8 Columns 3 Rows 1 Data Type

1 Change ZIP Code type to String
 2 Replace matches of 'NO USAR' from Company Name with ''
 3 Replace matches of 'no usar' from Company Name with ''
 4 Replace matches of 'No usar' from Company Name with ''

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

Columns Rows

RBC	Country	RBC	Company Name
165 Categories	IT	45,031 Categories	P.M. SERVICE SPA - NON USA
AE			non_usare*****MERKUR-Ueberseehandel GmbH

Clear all filters

Rows

RBC	Country	RBC	Company Name
165 Categories	IT	45,031 Categories	GLOBAL ENERGY S.R.L. (DA N)
IT			NON UTILIZZARE -- CIEL IMPIANTI S.R.L.
AE			TOTAL SOLAR - NON UTILIZZARE
IT			****non utilizzare***JSC Hydroenergy
IT			*** non utilizzareCAMERA DI COMMERCIO UFFICIALE SPAGNOLA IN

Clear all filters

New Step Recipe X

- 1 Change ZIP Code type to String
- 2 Replace matches of 'NO USAR' from Company Name with ''
- 3 Replace matches of 'no user' from Company Name with ''
- 4 Replace matches of 'No usar' from Company Name with ''
- 5 Replace matches of 'NO UTILIZAR' from Company Name with ''

8 Columns 2 Rows 1 Data Type

VANILLA FLOW > companies > Full Data

Columns Rows

RBC	Country	RBC	Company Name
165 Categories	IT	45,031 Categories	GLOBAL ENERGY S.R.L. (DA N)
IT			NON UTILIZZARE -- CIEL IMPIANTI S.R.L.
AE			TOTAL SOLAR - NON UTILIZZARE
IT			****non utilizzare***JSC Hydroenergy
IT			*** non utilizzareCAMERA DI COMMERCIO UFFICIALE SPAGNOLA IN

Clear all filters

Rows

RBC	Country	RBC	Company Name
165 Categories	IT	45,031 Categories	GLOBAL ENERGY S.R.L. (DA N)
IT			NON UTILIZZARE -- CIEL IMPIANTI S.R.L.
AE			TOTAL SOLAR - NON UTILIZZARE
IT			****non utilizzare***JSC Hydroenergy
IT			*** non utilizzareCAMERA DI COMMERCIO UFFICIALE SPAGNOLA IN

Clear all filters

New Step Recipe X

- 1 Change ZIP Code type to String
- 2 Replace matches of 'NO USAR' from Company Name with ''
- 3 Replace matches of 'no user' from Company Name with ''
- 4 Replace matches of 'No usar' from Company Name with ''
- 5 Replace matches of 'NO UTILIZAR' from Company Name with ''
- 6 Replace matches of 'non usare' from Company Name with ''
- 7 Replace matches of 'NON USARE' from Company Name with ''

8 Columns 5 Rows 1 Data Type

Now we standardise the company names.

Inconsistent company names demo Standardising values in a text column Celia Muriel – celiamuriel.com



The screenshot shows a data analysis interface with various tools and panels. On the left, there's a sidebar with icons for file operations, data types, and search. The main area has a toolbar at the top with icons for selection, filtering, and sorting. Below the toolbar is a table with two columns: 'Country' and 'Company Name'. A context menu is open over the 'Company Name' column, with the 'Standardize...' option highlighted. To the right of the table, there are several panels: 'Details' (with tabs for 'ABC Company Name', 'Quality', 'Unique Values', 'Patterns', and 'Suggestions'), a 'Quality' panel showing metrics like Valid (49781), Mismatched (0), and Missing (7), a 'Unique Values' panel listing categories like NULL, Koop-Brinkmann GmbH, PERSONA FISICA, and GRUPPO SONNEDIX, a 'Patterns' panel showing regular expressions and their counts, and a 'Suggestions' panel.

After clicking on “Standardize”, the company names are clustered by similar strings (the values have characters in common). This is a fuzzy matching method, as the company names within a cluster are not necessarily identical.

If all values in a cluster should have the same name, we select the cluster.

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

Row count ▾ Source value New value 6 values · 9 rows

2	Ingeteam S.R.L.	Ingeteam S.R.L.
2	INGETEAM s.r.l.	INGETEAM s.r.l.
2	INGETEAM S.R.L.	INGETEAM S.R.L.
1	"INGETEAM S.R.L."	"INGETEAM S.R.L."
1	***INGETEAM S.R.L.	***INGETEAM S.R.L.
1	Ingeteam s.r.l.	Ingeteam s.r.l.

4 values · 4 rows

1	"GENERAL ELECTRIC INTERNATIONAL, Inc."	"GENERAL ELECTRIC INTERNATIONAL, Inc."
1	"GENERAL ELECTRIC INTERNATIONAL, Inc"	"GENERAL ELECTRIC INTERNATIONAL, Inc"
1	GENERAL ELECTRIC INTERNATIONAL INC	GENERAL ELECTRIC INTERNATIONAL INC
1	GENERAL ELECTRIC INTERNATIONAL INC.	GENERAL ELECTRIC INTERNATIONAL INC.

4 values · 4 rows

1	Gamesa Wind GmbH	Gamesa Wind GmbH
1	Gamesa Wind GMBH	Gamesa Wind GMBH
1	GAMESA WIND GmbH	GAMESA WIND GmbH
1	GAMESA WIND GMBH	GAMESA WIND GMBH

4 values · 8 rows

4	INGETEAM Gmbh	INGETEAM Gmbh
2	INGETEAM GMBH	INGETEAM GMBH

1,010 clusters 45,031 unique source values 49,781 rows

Select a row to edit.

Standardize

Summary

Source column Company Name
 Unique new values 45031
 Source values updated 0 / 45031 (0.00%)
 Rows updated 0 / 49781 (0.00%)

Cancel Add to Recipe

We type the new value all occurrences in the cluster should have. Then we click on “Clustering options”.

VANILLA FLOW > companies > Full Data

Row count ▾ Source value New value 6 values · 9 rows

<input checked="" type="checkbox"/>	Ingeteam S.R.L.	Ingeteam S.R.L.
<input checked="" type="checkbox"/>	INGETEAM s.r.l.	Ingeteam S.R.L.
<input checked="" type="checkbox"/>	INGETEAM S.R.L.	Ingeteam S.R.L.
<input checked="" type="checkbox"/>	"INGETEAM S.R.L."	Ingeteam S.R.L.
<input checked="" type="checkbox"/>	***INGETEAM S.R.L.	Ingeteam S.R.L.
<input checked="" type="checkbox"/>	Ingeteam s.r.l.	Ingeteam S.R.L.

4 values · 4 rows

<input type="checkbox"/>	"GENERAL ELECTRIC INTERNATIONAL, Inc."	"GENERAL ELECTRIC INTERNATIONAL, Inc."
<input type="checkbox"/>	"GENERAL ELECTRIC INTERNATIONAL, Inc"	"GENERAL ELECTRIC INTERNATIONAL, Inc"
<input type="checkbox"/>	GENERAL ELECTRIC INTERNATIONAL INC	GENERAL ELECTRIC INTERNATIONAL INC
<input type="checkbox"/>	GENERAL ELECTRIC INTERNATIONAL INC.	GENERAL ELECTRIC INTERNATIONAL INC.

4 values · 4 rows

<input type="checkbox"/>	Gamesa Wind GmbH	Gamesa Wind GmbH
<input type="checkbox"/>	Gamesa Wind GMBH	Gamesa Wind GMBH
<input type="checkbox"/>	GAMESA WIND GmbH	GAMESA WIND GmbH
<input type="checkbox"/>	GAMESA WIND GMBH	GAMESA WIND GMBH

4 values · 8 rows

<input type="checkbox"/>	INGETEAM Gmbh	INGETEAM Gmbh
<input type="checkbox"/>	INGETEAM GMBH	INGETEAM GMBH

New value Revert to source Apply

Source value Multiple values

Row count 9

Summary

Source column Company Name
 Unique new values 45031
 Source values updated 0 / 45031 (0.00%)
 Rows updated 0 / 49781 (0.00%)

Cancel Add to Recipe



We confirm we were clustering on similar strings.

VANILLA FLOW > companies > Full Data

Row count > Source value New value

Source value	New value	Count
2 Ingeteam S.R.L.	Ingeteam S.R.L.	6 values · 9 rows
2 INGETEAM s.r.l.	Ingeteam S.R.L.	
2 INGETEAM-S.R.L.	Ingeteam S.R.L.	
1 "INGETEAM, S.R.L."	Ingeteam S.R.L.	
1 ***INGETEAM S.R.L.	Ingeteam S.R.L.	
1 Ingeteam s.r.l.	Ingeteam S.R.L.	
1 "GENERAL-ELECTRIC INTERNATIONAL, Inc"	"GENERAL-ELECTRIC INTERNATIONAL, Inc"	4 values · 4 rows
1 "GENERAL-ELECTRIC INTERNATIONAL, Inc"	"GENERAL-ELECTRIC INTERNATIONAL, Inc"	
1 GENERAL ELECTRIC INTERNATIONAL INC	GENERAL ELECTRIC INTERNATIONAL INC	
1 GENERAL ELECTRIC INTERNATIONAL INC	GENERAL ELECTRIC INTERNATIONAL INC	
1 Gamesa-Wind GmbH	Gamesa Wind GmbH	4 values · 4 rows
1 Gamesa-Wind GMBH	Gamesa Wind GMBH	
1 GAMESA-WIND GmbH	GAMESA WIND GmbH	
1 GAMESA-WIND GMBH	GAMESA WIND GMBH	
4 INGETEAM-Gmbh	INGETEAM-Gmbh	4 values · 8 rows
2 INGETEAM-GMBH	INGETEAM-GMBH	
2 INGETEAM-GmbH	INGETEAM-GmbH	
2 INGETEAM-GMBH	INGETEAM-GMBH	

1,010 clusters 45,031 unique source values 49,781 rows 6 selected (9 rows)

Clustering options

Clustering method

- None
- Do not cluster values
- Similar strings** Cluster values that have characters in common
- Pronunciation
- Cluster values that sound alike

Model

Fingerprint Ngram

Now we cluster on similar pronunciation, that's to say, the values sound alike. We select uncheck values in the new cluster if they should have the same value as the cluster we have already selected with the similar string.

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

Row count	Source value	New value
1	YC2·ENERJY	YC2·ENERJY
2	Ingeteam S.R.L.	Ingeteam S.R.L.
2	INGETEAM·s.r.l.	Ingeteam S.R.L.
2	INGETEAM S.R.L.	Ingeteam S.R.L.
2	INGETEAM SRL	INGETEAM SRL
1	"INGETEAM,-S.R.L.-"	Ingeteam S.R.L.
1	***INGETEAM S.R.L.	Ingeteam S.R.L.
1	Ingeteam s.r.l.	Ingeteam S.R.L.
2	"ESF·SPANIEN·0301,-S.L.-"	"ESF·SPANIEN·0301,-S.L.-"
2	"ESF·SPANIEN·0303,-S.L.-"	"ESF·SPANIEN·0303,-S.L.-"
2	ESF·Spanien 0302 S.L.U.	ESF·Spanien 0302 S.L.U.
2	ESF·Spanien 0306 S.L.U.	ESF·Spanien 0306 S.L.U.
2	ESF·Spanien 0308 S.L.U.	ESF·Spanien 0308 S.L.U.
2	ESF·Spanien 0311 S.L.U.	ESF·Spanien 0311 S.L.U.
1	ESF·SPANIEN 0424 SL	ESF·SPANIEN 0424 SL
2	IS ENERGY SRL	IS ENERGY SRL
1	"EOSA·ENERGIA,-S.R.L.-"	"EOSA·ENERGIA,-S.R.L.-"
1	IS ENERGY SRL	IS ENERGY SRL
2,410 clusters	45,031 unique source values	49,781 rows
6 selected (9 rows)		

Clustering options

Clustering method

- None
- Do not cluster values
- Similar strings
- Cluster values that have characters in common
- Pronunciation**
- Cluster values that sound alike

VANILLA FLOW > companies > Full Data

Row count	Source value	New value
1	YC2·ENERJY	YC2·ENERJY
2	Ingeteam S.R.L.	Ingeteam S.R.L.
2	INGETEAM·s.r.l.	Ingeteam S.R.L.
2	INGETEAM S.R.L.	Ingeteam S.R.L.
2	INGETEAM SRL	Ingeteam S.R.L.
1	"INGETEAM,-S.R.L.-"	Ingeteam S.R.L.
1	***INGETEAM S.R.L.	Ingeteam S.R.L.
1	Ingeteam s.r.l.	Ingeteam S.R.L.
2	"ESF·SPANIEN·0301,-S.L.-"	"ESF·SPANIEN·0301,-S.L.-"
2	"ESF·SPANIEN·0303,-S.L.-"	"ESF·SPANIEN·0303,-S.L.-"
2	ESF·Spanien 0302 S.L.U.	ESF·Spanien 0302 S.L.U.
2	ESF·Spanien 0306 S.L.U.	ESF·Spanien 0306 S.L.U.
2	ESF·Spanien 0308 S.L.U.	ESF·Spanien 0308 S.L.U.
2	ESF·Spanien 0311 S.L.U.	ESF·Spanien 0311 S.L.U.
1	ESF·SPANIEN 0424 SL	ESF·SPANIEN 0424 SL
2	IS ENERGY SRL	IS ENERGY SRL
1	"EOSA·ENERGIA,-S.R.L.-"	"EOSA·ENERGIA,-S.R.L.-"
1	IS ENERGY SRL	IS ENERGY SRL
2,410 clusters	45,031 unique source values	49,781 rows
7 selected (11 rows)		

Clustering options

Clustering method

- None
- Do not cluster values
- Similar strings
- Cluster values that have characters in common
- Pronunciation**
- Cluster values that sound alike

We apply the new value to all the occurrences we selected (similar string or pronunciation).
 Bear in mind that the Double Metaphone algorithm - used to cluster values with similar

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



phonetics - fails when there is a space among the characters. You should review all possible values for the company names in the CSV file used for this demo.

VANILLA FLOW > companies > Full Data

Clustering options Search values... (/)

Row count	Source value	New value
1	YC2·ENERJY	YC2·ENERJY
7	2 Ingeteam·S.R.L. 2 INGETEAM s.r.l. 2 INGETEAM S.R.L. 2 INGETEAM·SRL 1 "INGETEAM, S.R.L." 1 ***INGETEAM S.R.L. 1 Ingeteam·s.r.l.	Ingeteam·S.R.L. Ingeteam·S.R.L. Ingeteam·S.R.L. Ingeteam·S.R.L. Ingeteam·S.R.L. Ingeteam·S.R.L. Ingeteam·S.R.L.
13	2 "ESF·SPANIEN 0301, S.L." 2 "ESF·SPANIEN 0303, S.L." 2 ESF·Spanien 0302 S.L.U. 2 ESF·Spanien 0306 S.L.U. 2 ESF·Spanien 0308 S.L.U. 2 ESF·Spanien 0311 S.L.U. 1 ESF·SPANIEN 0424 SL	"ESF·SPANIEN 0301, S.L." "ESF·SPANIEN 0303, S.L." ESF·Spanien 0302 S.L.U. ESF·Spanien 0306 S.L.U. ESF·Spanien 0308 S.L.U. ESF·Spanien 0311 S.L.U. ESF·SPANIEN 0424 SL
8	2 IS·ENERGY·SRL 1 "EOSA·ENERGIA, S.R.L." 1 IS·ENERGY·SRL	IS·ENERGY·SRL "EOSA·ENERGIA, S.R.L." IS·ENERGY·SRL
45,031	45,031 unique source values	49,781 rows

Standardize

New value: Ingeteam S.R.L.

Source value: Multiple values

Row count: 11

Summary

Source column Company Name
 Unique new values 45031
 Source values updated 0 / 45031 (0.00%)
 Rows updated 0 / 49781 (0.00%)

VANILLA FLOW > companies > Full Data

Clustering options Search values... (/)

Row count	Source value	New value
1	YC2·ENERJY	YC2·ENERJY
7	2 Ingeteam·S.R.L. 2 INGETEAM s.r.l. 2 INGETEAM S.R.L. 2 INGETEAM·SRL 1 "INGETEAM, S.R.L." 1 ***INGETEAM S.R.L. 1 Ingeteam·s.r.l.	Ingeteam·S.R.L. Ingeteam·S.R.L. Ingeteam·S.R.L. Ingeteam·S.R.L. Ingeteam·S.R.L. Ingeteam·S.R.L. Ingeteam·S.R.L.
13	2 "ESF·SPANIEN 0301, S.L." 2 "ESF·SPANIEN 0303, S.L." 2 ESF·Spanien 0302 S.L.U. 2 ESF·Spanien 0306 S.L.U. 2 ESF·Spanien 0308 S.L.U. 2 ESF·Spanien 0311 S.L.U. 1 ESF·SPANIEN 0424 SL	"ESF·SPANIEN 0301, S.L." "ESF·SPANIEN 0303, S.L." ESF·Spanien 0302 S.L.U. ESF·Spanien 0306 S.L.U. ESF·Spanien 0308 S.L.U. ESF·Spanien 0311 S.L.U. ESF·SPANIEN 0424 SL
8	2 IS·ENERGY·SRL 1 "EOSA·ENERGIA, S.R.L." 1 IS·ENERGY·SRL	IS·ENERGY·SRL "EOSA·ENERGIA, S.R.L." IS·ENERGY·SRL
45,031	45,031 unique source values	49,781 rows

Standardize

Select a row to edit.

Summary

Source column Company Name
 Unique new values 45025
 Source values updated 6 / 45031 (0.01%)
 Rows updated 9 / 49781 (0.02%)



There is also a wizard button which will automatically standardise all values within each cluster to a suggested value based on occurrence frequency. Use it if you are fairly confident Cloud Dataprep will easily detect all changes which need to be done. Then you can review the changes before adding to the recipe.

VANILLA FLOW > companies > Full Data

Clustering options Search values... (/)

Row count Source value New value

1	YC2-ENERJY	YC2-ENERJY
7 values - 11 rows		
2	Ingeteam-S.R.L.	Ingeteam-S.R.L.
2	INGETEAM-s.r.l.	Ingeteam-S.R.L.
2	INGETEAM-S.R.L.	Ingeteam-S.R.L.
2	INGETEAM-SRL	Ingeteam-S.R.L.
1	"INGETEAM-S.R.L."	Ingeteam-S.R.L.
1	***INGETEAM-S.R.L.	Ingeteam-S.R.L.
1	Ingeteam-s.r.l.	Ingeteam-S.R.L.
7 values - 13 rows		
2	"ESF-SPANIEN-0301, S.L."	"ESF-SPANIEN-0301, S.L."
2	"ESF-SPANIEN-0303, S.L."	"ESF-SPANIEN-0303, S.L."
2	ESF-Spanien-0302 S.L.U.	ESF-Spanien-0302 S.L.U.
2	ESF-Spanien-0306 S.L.U.	ESF-Spanien-0306 S.L.U.
2	ESF-Spanien-0308 S.L.U.	ESF-Spanien-0308 S.L.U.
2	ESF-Spanien-0311 S.L.U.	ESF-Spanien-0311 S.L.U.
1	ESF-SPANIEN-0424 SL	ESF-SPANIEN-0424 SL
7 values - 8 rows		
2	IS ENERGY-SRL	IS ENERGY-SRL
1	"EOSA-ENERGIA-S.R.L."	"EOSA-ENERGIA-S.R.L."
1	EOA-ENERGIA-SRL	EOA-ENERGIA-SRL
2,410 clusters	45,031 unique source values	49,781 rows

Select a row to edit.

Summary

Source column	Company Name
Unique new values	45025
Source values updated	6 / 45031 (0.01%)
Rows updated	9 / 49781 (0.02%)

Cancel Add to Recipe

Once you are pleased with all values in the CSV, click on “Save to recipe”.

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

Standardize

Select a row to edit.

Row count	Source value	New value
2	Ingeteam S.R.L.	Ingeteam S.R.L.
2	INGETEAM s.r.l.	Ingeteam S.R.L.
2	INGETEAM S.R.L.	Ingeteam S.R.L.
1	"INGETEAM, S.R.L."	Ingeteam S.R.L.
1	***INGETEAM S.R.L.	Ingeteam S.R.L.
1	Ingeteam s.r.l.	Ingeteam S.R.L.

Row count	Source value	New value
1	"GENERAL ELECTRIC INTERNATIONAL, Inc."	General Electric International Inc.
1	"GENERAL ELECTRIC INTERNATIONAL, Inc"	General Electric International Inc.
1	GENERAL ELECTRIC INTERNATIONAL INC	General Electric International Inc.
1	GENERAL ELECTRIC INTERNATIONAL INC.	General Electric International Inc.

Row count	Source value	New value
1	Gamesa Wind GmbH	Gamesa Wind GmbH
1	Gamesa Wind GMBH	Gamesa Wind GmbH
1	GAMESA WIND GmbH	Gamesa Wind GmbH
1	GAMESA WIND GMBH	Gamesa Wind GmbH

Row count	Source value	New value
4	INGETEAM GmbH	Ingeteam GmbH
2	INGETEAM GMBH	Ingeteam GmbH

Summary

Source column	Company Name
Unique new values	44686
Source values updated	368 / 45031 (0.82%)
Rows updated	425 / 49781 (0.85%)

Add to Recipe

VANILLA FLOW > companies > Full Data

Recipe

1 Change ZIP Code type to String
 2 Replace matches of 'NO USAR' from Company Name with ''
 3 Replace matches of 'no user' from Company Name with ''
 4 Replace matches of 'No user' from Company Name with ''
 5 Replace matches of 'NO UTILIZAR' from Company Name with ''
 6 Replace matches of 'non usare' from Company Name with ''
 7 Replace matches of 'NON USARE' from Company Name with ''
 8 Replace matches of 'non utilizzare' from Company Name with ''
 9 Replace matches of 'NON UTILIZZARE' from Company Name with ''
 10 Standardize Company Name

We remove the leading and trailing quotes in the company names.

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

Run Job

RBC Country RBC Company Name RBC Company ID 1 RBC

165 Categories 44,686 Categories

ES SANJOSE · TECNOLOGIAS SA
 ES "ALCAZAREN SOSTENIBLE, S.L."
 KR "SAMSUNG Heavy Industries Co., LTD."
 DE Avalon Solar & Wind GmbH
 DE HAWI Energietechnik AG
 DE "SONNENZINS SOLAR, Gmbh."
 ES "INST. FOTOVOLTAICAS VOLFTER, S.L."
 ES "ELECTRICIDAD FRAN RENOVABLES, SLU"
 ES "SERVIMA LEVANTE, SL"
 ES ENERGIAS ALTERNATIVAS ARNEDO
 CN "HUAYI WIND ENERGY CO., LTD ZHEJIANG HEWIND
 ES HERNANDEZ RUIZ ANDRES
 ES "DELEGADA TECNICA, S.L."
 FR Systovi
 ES TERREROS SERRANO ANTONIO
 FR VANITEX - SOLARENT
 FR Ecomotiv
 FR Dja Stock Energie
 FR Solea
 ES ADJ DITEC MALAGA SLL

8 Columns 49,788 Rows 1 Data Type

Format Convert to UPPERCASE
 Calculate Convert to lowercase
 Create column from examples Convert to Proper Case
 Group by Trim leading and trailing whitespace
 Pivot Trim leading and trailing quotes
 Restructure Remove whitespace
 Filter rows Remove symbols
 Replace Remove accents
 Standardize...
 Extract
 Split column
 Column Details
 Show related Steps in Recipe

VANILLA FLOW > companies > Full Data

Run Job

Source to be dropped Preview

RBC Country RBC Company Name RBC Company Name

165 Categories 44,686 Categories

ES SANJOSE · TECNOLOGIAS SA
 ES "ALCAZAREN SOSTENIBLE, S.L."
 KR "SAMSUNG Heavy Industries Co., LTD."
 DE Avalon Solar & Wind GmbH
 DE HAWI Energietechnik AG
 DE "SONNENZINS SOLAR, Gmbh."
 ES "INST. FOTOVOLTAICAS VOLFTER, S.L."
 ES "ELECTRICIDAD FRAN RENOVABLES, SLU"
 ES "SERVIMA LEVANTE, SL"
 ES ENERGIAS ALTERNATIVAS ARNEDO
 CN "HUAYI WIND ENERGY CO., LTD ZHEJIANG HEWIND
 ES HERNANDEZ RUIZ ANDRES
 ES "DELEGADA TECNICA, S.L."
 FR Systovi
 ES TERREROS SERRANO ANTONIO
 FR VANITEX - SOLARENT
 FR Ecomotiv
 FR Dja Stock Energie
 FR Solea
 ES ADJ DITEC MALAGA SLL

Text format

Columns required
 Multiple
 RBC Company Name
 Format required
 Trim leading and trailing quotes
 Remove quotes found at the beginning and end of the text

Add

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

RBC	Country	RBC	Company Name	RBC	Company ID 1	RBC
	165 Categories		44,686 Categories		22,191 Categories	28 Categories
-	ES	-	SANJOSE TECNOLOGIAS SA	-	A36409910	
-	ES	-	ALCAZAREN SOSTENIBLE, S.L.	-	B37468170	
-	KR	-	SAMSUNG Heavy Industries Co., LTD.	-	612850343	
-	DE	-	Avalon Solar & Wind GmbH	-	204689356	
-	DE	-	HAWI Energietechnik AG	-	DE813293213	
-	DE	-	SONNENZINS SOLAR, GmbH.	-	DE244407327	
-	ES	-	INST. FOTOVOLTAICAS VOLFTER, S.L.	-	B36993558	
-	ES	-	ELECTRICIDAD FRAN RENOVABLES, SLU	-	B54358254	
-	ES	-	SERVIMA LEVANTE, SL	-	B97822388	
-	ES	-	ENERGIAS ALTERNATIVAS ARNEDO	-	B26382143	
-	CN	-	HUAYI WIND ENERGY CO., LTD ZHEJIANG HEWIND C	-	3300736897797	
-	ES	-	HERNANDEZ RUIZ ANDRES	-	704137570	
-	ES	-	DELEGADA TECNICA, S.L.	-	B46822711	
-	FR	-	Systovi	-	509496998	
-	ES	-	TERREROS SERRANO ANTONIO	-	04569857X	
-	FR	-	VANITEX - SOLARENT	-	514555416	
-	FR	-	Ecomotiv	-	511253346	
-	FR	-	Dja Stock Energie	-	411925118	
-	FR	-	Solea	-	497551994	
-	ES	-	ADJ-DITEC MALAGA SLL	-	B93010981	

8 Columns 49,788 Rows 1 Data Type

→ 11 Trim quotes from Company Name

We remove the leading and trailing white spaces in the company names.

VANILLA FLOW > companies > Full Data

RBC	Country	RBC	Company Name	RBC	Company ID 1	RBC
	165 Categories		44,686 Categories		22,191 Categories	28 Categories
-	ES	-	SANJOSE TECNOLOGIAS SA	-	A36409910	
-	ES	-	ALCAZAREN SOSTENIBLE, S.L.	-	B37468170	
-	KR	-	SAMSUNG Heavy Industries Co., LTD.	-	612850343	
-	DE	-	Avalon Solar & Wind GmbH	-	204689356	
-	DE	-	HAWI Energietechnik AG	-	DE813293213	
-	DE	-	SONNENZINS SOLAR, GmbH.	-	DE244407327	
-	ES	-	INST. FOTOVOLTAICAS VOLFTER, S.L.	-	B36993558	
-	ES	-	ELECTRICIDAD FRAN RENOVABLES, SLU	-	B54358254	
-	ES	-	SERVIMA LEVANTE, SL	-	B97822388	
-	ES	-	ENERGIAS ALTERNATIVAS ARNEDO	-	B26382143	
-	CN	-	HUAYI WIND ENERGY CO., LTD ZHEJIANG HEWIND C	-	3300736897797	
-	ES	-	HERNANDEZ RUIZ ANDRES	-	704137570	
-	ES	-	DELEGADA TECNICA, S.L.	-	B46822711	
-	FR	-	Systovi	-	509496998	
-	ES	-	TERREROS SERRANO ANTONIO	-	04569857X	
-	FR	-	VANITEX - SOLARENT	-	514555416	
-	FR	-	Ecomotiv	-	511253346	
-	FR	-	Dja Stock Energie	-	411925118	
-	FR	-	Solea	-	497551994	
-	ES	-	ADJ-DITEC MALAGA SLL	-	B93010981	

8 Columns 49,788 Rows 1 Data Type

→ 11 Trim quotes from Company Name

Context menu for the Company Name column:

- Rename
- Change type
- Move
- Hide
- Format** (highlighted)
- Calculate
- Create column from examples
- Group by
- Trim leading and trailing whitespace** (highlighted)
- Pivot
- Restructure
- Filter rows
- Replace
- Standardize...
- Extract
- Split column
- Column Details
- Show related Steps in Recipe

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

Source	to be dropped	Preview			
RBC	Country	RBC	Company Name	RBC	Company Name
165 Categories	44,686 Categories	44,449 Categories	SANJOSE TECNOLOGIAS SA	SANJOSE TECNOLOGIAS SA	
- ES	ALCAZAREN SOSTENIBLE, S.L.		ALCAZAREN SOSTENIBLE, S.L.	ALCAZAREN SOSTENIBLE, S.L.	
- KR	SAMSUNG Heavy Industries Co., LTD..		SAMSUNG Heavy Industries Co., LTD..	SAMSUNG Heavy Industries Co., LTD..	
- DE	Avalon Solar & Wind GmbH		Avalon Solar & Wind GmbH	Avalon Solar & Wind GmbH	
- DE	HAWI Energietechink AG		HAWI Energietechink AG	HAWI Energietechink AG	
- DE	SONNENZINS SOLAR, GmbH..		SONNENZINS SOLAR, GmbH..	SONNENZINS SOLAR, GmbH..	
- ES	INST. FOTOVOLTAICAS VOLFTER, S.L.		INST. FOTOVOLTAICAS VOLFTER, S.L.	INST. FOTOVOLTAICAS VOLFTER, S.L.	
- ES	ELECTRICIDAD FRAN RENOVABLES, SLU		ELECTRICIDAD FRAN RENOVABLES, SLU	ELECTRICIDAD FRAN RENOVABLES, SLU	
- ES	SERVIMA LEVANTE, SL		SERVIMA LEVANTE, SL	SERVIMA LEVANTE, SL	
- ES	ENERGIAS ALTERNATIVAS ARNEDO		ENERGIAS ALTERNATIVAS ARNEDO	ENERGIAS ALTERNATIVAS ARNEDO	
- CN	HUAYI WIND ENERGY CO., LTD ZHEJIANG HEWIND		HUAYI WIND ENERGY CO., LTD ZHEJIANG HEWIND	HUAYI WIND ENERGY CO., LTD ZHEJIANG HEWIND	
- ES	HERNANDEZ RUIZ ANDRES		HERNANDEZ RUIZ ANDRES	HERNANDEZ RUIZ ANDRES	
- ES	DELEGADA TECNICA, S.L..		DELEGADA TECNICA, S.L..	DELEGADA TECNICA, S.L..	
- FR	Systovi		Systovi	Systovi	
- ES	TERREROS SERRANO ANTONIO		TERREROS SERRANO ANTONIO	TERREROS SERRANO ANTONIO	
- FR	VANITEX - SOLARENT		VANITEX - SOLARENT	VANITEX - SOLARENT	
- FR	Ecomotiv		Ecomotiv	Ecomotiv	
- FR	Dja Stock Energie		Dja Stock Energie	Dja Stock Energie	
- FR	Solea		Solea	Solea	
- ES	ADJ-DITEC MALAGA-SLL		ADJ-DITEC MALAGA-SLL	ADJ-DITEC MALAGA-SLL	

9 Columns 49,788 Rows 1 Data Type

Text format required

Columns required

RBC Company Name

Format required

Trim leading and trailing whitespace

Remove all whitespaces found at the beginning and end of the text

Add

VANILLA FLOW > companies > Full Data

RBC	Country	RBC	Company Name	RBC	Company ID 1	RBC
165 Categories	44,449 Categories	22,191 Categories	28 Categories			
- ES	SANJOSE TECNOLOGIAS SA	A36409910				
- ES	ALCAZAREN SOSTENIBLE, S.L.	B37468170				
- KR	SAMSUNG Heavy Industries Co., LTD..	6128500343				
- DE	Avalon Solar & Wind GmbH	204689356				
- DE	HAWI Energietechink AG	DE813293213				
- DE	SONNENZINS SOLAR, GmbH..	DE244407327				
- ES	INST. FOTOVOLTAICAS VOLFTER, S.L.	B36993558				
- ES	ELECTRICIDAD FRAN RENOVABLES, SLU	B54358254				
- ES	SERVIMA LEVANTE, SL	B97822388				
- ES	ENERGIAS ALTERNATIVAS ARNEDO	B26382143				
- CN	HUAYI WIND ENERGY CO., LTD ZHEJIANG HEWIND	3300736897797				
- ES	HERNANDEZ RUIZ ANDRES	70413757Q				
- ES	DELEGADA TECNICA, S.L..	B46022711				
- FR	Systovi	509496998				
- ES	TERREROS SERRANO ANTONIO	04569857X				
- FR	VANITEX - SOLARENT	514555416				
- FR	Ecomotiv	511253346				
- FR	Dja Stock Energie	411925118				
- FR	Solea	497551994				
- ES	ADJ-DITEC MALAGA-SLL	B93010981				

New Step Recipe

3 Replace matches of 'no usar' from Company Name with "

4 Replace matches of 'No usar' from Company Name with "

5 Replace matches of 'NO UTILIZAR' from Company Name with "

6 Replace matches of 'non usare' from Company Name with "

7 Replace matches of 'NON USARE' from Company Name with "

8 Replace matches of 'non utilizzare' from Company Name with "

9 Replace matches of 'NON UTILIZZARE' from Company Name with "

10 Standardize Company Name

11 Trim quotes from Company Name

12 Trim whitespace from Company Name

We change the format in the Town field to have the first letter in uppercase and the rest in lowercase.

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

Company ID 3 RBC Company ID 4 RBC Town RBC ZIP Code

20,050 Categories 16,469 Categories

Madrid BARBALOS DE HUEBRA (SALAMANCA) "445-330 Hwsung City, Gyeonggi" Biebergemünd Eggenfelden Essenbach-Altheim VIGO ELCHE Paterna ARNEDO Zhejiang OROPESA (Toledo) Valencia Saint Herblain MADRID Paris Reventin Limoges Mauguio MALAGA

8 Columns 49,788 Rows 1 Data Type

Format Convert to UPPERCASE
Convert to lowercase
Convert to Proper Case
Calculate Create column from examples
Group by Pivot Restructure
Filter rows Replace Standardize...
Remove whitespace Remove symbols Remove accents

VANILLA FLOW > companies > Full Data

Source to be dropped Preview

3 RBC Company ID 4 RBC Town RBC Town

20,050 Categories 16,469 Categories 14,827 Categories

Madrid BARBALOS DE HUEBRA (SALAMANCA) "445-330 Hwsung City, Gyeonggi" Biebergemünd Eggenfelden Essenbach-Altheim VIGO ELCHE Paterna ARNEDO Zhejiang OROPESA (Toledo) Valencia Saint Herblain MADRID Paris Reventin Limoges Mauguio MALAGA

9 Columns 49,788 Rows 1 Data Type

Text format

Columns required
Multiple
RBC Town

Format required
Convert to Proper Case
Convert text in column to ProperCase

Add

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

Run Job

Company ID 3 RBC Company ID 4 RBC Town ZIP Code

20,050 Categories 14,827 Categories 12,026 Categories

Madrid 28760
 Barbalo De Huebra (Salamanca) 37455
 "445-330 Hwsung City, Gyeonggi-do 445-330
 Biebergemünd 63599
 Eggentalen 84307
 Essenbach-Altheim 84051
 Vigo 36214
 Elche 3205
 Paterna 46980
 Arnedo 26580
 Zhejiang 325600
 Oropesa (Toledo) 45560
 Valencia 46809
 Saint-Herblain 44806
 Madrid 28845
 Paris 75002
 Reventin 38121
 Limoges 87100
 Mauguio 34130
 Malaga 29806

8 Columns 49,788 Rows 1 Data Type

14 Convert text in Town to Propercase

We remove accents.

VANILLA FLOW > companies > Full Data

Run Job

Company ID 3 RBC Company ID 4 RBC Town ZIP Code

20,050 Categories 14,827 Categories

Madrid
 Barbalo De Huebra (Salamanca)
 "445-330 Hwsung City, Gyeonggi-do 445-330
 Biebergemünd
 Eggentalen
 Essenbach-Altheim
 Vigo
 Elche
 Paterna
 Arnedo
 Zhejiang
 Oropesa (Toledo)
 Valencia
 Saint-Herblain
 Madrid
 Paris
 Reventin
 Limoges
 Mauguio
 Malaga

Rename
 Change type
 Move
 Hide
 Format
 Calculate
 Create column from examples
 Group by
 Pivot
 Restructure
 Filter rows
 Replace
 Standardize...
 Extract
 Split column
 Column Details
 Show related Steps in Recipe

Convert to UPPERCASE
 Convert to lowercase
 Convert to Proper Case
 Trim leading and trailing whitespace
 Trim leading and trailing quotes
 Remove whitespace
 Remove symbols
 Remove accents

8 Columns 49,788 Rows 1 Data Type

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

Text format

Columns required
Multiple
RBC Town

Format required
Remove accents
Remove all accents from the text

Add

9 Columns 49,788 Rows 1 Data Type

VANILLA FLOW > companies > Full Data

New Step Recipe **X**

5 Replace matches of 'NO UTILIZAR' from Company Name with ''
 6 Replace matches of 'non usare' from Company Name with ''
 7 Replace matches of 'NON USARE' from Company Name with ''
 8 Replace matches of 'non utilizzare' from Company Name with ''
 9 Replace matches of 'NON UTILIZZARE' from Company Name with ''
 10 Standardize Company Name
 11 Trim quotes from Company Name
 12 Trim whitespace from Company Name
 13 Rename Country1 to 'Town'
 14 Convert text in Town to ProperCase
15 Remove accents from Town

8 Columns 49,788 Rows 1 Data Type

We remove symbols.

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

Company ID 3 RBC Company ID 4 RBC Town RBC ZIP Code

20,050 Categories 14,609 Categories

Madrid
 Barbalos De Huebra (Salamanca)
 "445-330 Hwsung City, Gyeonggi-Do"
 Biebergemund
 Eggenfelden
 Essenbach-Altheim
 Vigo
 Elche
 Paterna
 Arnedo
 Zhejiang
 Oropesa (Toledo)
 Valencia
 Saint-Herblain
 Madrid
 Paris
 Reventin
 Limoges
 Mauguio
 Malaga

Format Remove symbols

Convert to UPPERCASE
 Convert to lowercase
 Convert to Proper Case
 Trim leading and trailing whitespace
 Trim leading and trailing quotes
 Remove whitespace
 Remove symbols
 Remove accents

8 Columns 49,788 Rows 1 Data Type

VANILLA FLOW > companies > Full Data

Source to be dropped Preview

RBC Company ID 4 RBC Town RBC Town RBC

20,050 Categories 14,609 Categories 14,408 Categories

Madrid
 Barbalos De Huebra (Salamanca)
 "445-330 Hwsung City, Gyeonggi-Do"
 Biebergemund
 Eggenfelden
 Essenbach-Altheim
 Vigo
 Elche
 Paterna
 Arnedo
 Zhejiang
 Oropesa (Toledo)
 Valencia
 Saint-Herblain
 Madrid
 Paris
 Reventin
 Limoges
 Mauguio
 Malaga

Text format

Columns required
 Multiple ABC Town
 Format required
 Remove symbols
 Remove all non-alphanumeric characters from the text

Add

Cancel

9 Columns 49,788 Rows 1 Data Type

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

Company ID 3	RBC	Company ID 4	RBC	Town	RBC	ZIP Code
				Madrid		28760
				Barbalos De Huebra Salamanca		37455
				445330 Hwsung City GyeonggiDo		445-330
		DE204689356		Biebergemund		63599
		DE813293213		Eggenfelden		84307
		DE244407327		EssenbachAltheim		84051
		ESB36993558		Vigo		36214
				Elche		3205
				Paterna		46980
				Arnedo		26580
				Zhejiang		325600
				Oropesa Toledo		45560
				Valencia		46009
				Saint Herblain		44886
				Madrid		28045
				Paris		75002
				Reventin		38121
				Limoges		87100
				Mauguio		34130
				Malaga		29006

8 Columns 49,788 Rows 1 Data Type

16 Remove symbols from Town

We standardise the Town field with the Wizard.

VANILLA FLOW > companies > Full Data

Company ID 3	RBC	Company ID 4	RBC	Town	RBC	ZIP Code
				Madrid		28760
				Barbalos De Huebra Salamanca		37455
				445330 Hwsung City GyeonggiDo		445-330
		DE204689356		Biebergemund		63599
		DE813293213		Eggenfelden		84307
		DE244407327		EssenbachAltheim		84051
		ESB36993558		Vigo		36214
				Elche		3205
				Paterna		46980
				Arnedo		26580
				Zhejiang		325600
				Oropesa Toledo		45560
				Valencia		46009
				Saint Herblain		44886
				Madrid		28045
				Paris		75002
				Reventin		38121
				Limoges		87100
				Mauguio		34130
				Malaga		29006

8 Columns 49,788 Rows 1 Data Type

Standardize...

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

Standardize

Summary

Source column: Town
 Unique new values: 14408
 Source values updated: 0 / 14408 (0.00%)
 Rows updated: 0 / 37346 (0.00%)

Cancel **Add to Recipe**

Auto standardize changed 139 values in 132 clusters. Undo

Row count	Source value	New value	Count
9	Arezzo·Ar	Arezzo·Ar	4 values · 14 rows
3	Arezzo··Ar·	Arezzo··Ar·	
1	Arezzo··Ar	Arezzo··Ar	
1	Arezzo·Ar·	Arezzo·Ar·	
			3 values · 9 rows
5	Zamudio·Bizkaia	Zamudio Bizkaia	
3	Zamudio··Bizkaia	Zamudio··Bizkaia	
1	Zamudio··Bizkaia·	Zamudio··Bizkaia·	
			3 values · 6 rows
3	San·Giovanni·In·Persiceto··Bo·	San·Giovanni·In·Persiceto··Bo·	
2	San·Giovanni·In·Persiceto·Bo	San·Giovanni·In·Persiceto·Bo	
1	San·Giovanni··In·Persiceto·Bo	San·Giovanni··In·Persiceto·Bo	
			3 values · 11 rows
5	Civitanova·Marche··Mc·	Civitanova·Marche··Mc·	
5	Civitanova·Marche·Mc	Civitanova·Marche·Mc	
1	Civitanova·Marche···Mc·	Civitanova·Marche···Mc·	
			3 values · 15 rows
11	Las·Rozas··Madrid	Las·Rozas··Madrid	
221 clusters	14,408 unique source values	37,346 rows	

VANILLA FLOW > companies > Full Data

Standardize

Summary

Source column: Town
 Unique new values: –
 Source values updated: –
 Rows updated: –

Cancel **Add to Recipe**

Row count	Source value	New value	Count
9	Arezzo·Ar	Arezzo·Ar	4 values · 14 rows
3	Arezzo··Ar·	Arezzo··Ar·	
1	Arezzo··Ar	Arezzo··Ar	
1	Arezzo·Ar·	Arezzo·Ar·	
			3 values · 9 rows
5	Zamudio·Bizkaia	Zamudio Bizkaia	
3	Zamudio··Bizkaia	Zamudio··Bizkaia	
1	Zamudio··Bizkaia·	Zamudio··Bizkaia·	
			3 values · 6 rows
3	San·Giovanni·In·Persiceto··Bo·	San·Giovanni·In·Persiceto··Bo·	
2	San·Giovanni·In·Persiceto·Bo	San·Giovanni·In·Persiceto·Bo	
1	San·Giovanni··In·Persiceto·Bo	San·Giovanni··In·Persiceto·Bo	
			3 values · 11 rows
5	Civitanova·Marche··Mc·	Civitanova·Marche··Mc·	
5	Civitanova·Marche·Mc	Civitanova·Marche·Mc	
1	Civitanova·Marche···Mc·	Civitanova·Marche···Mc·	
			3 values · 15 rows
11	Las·Rozas··Madrid	Las·Rozas··Madrid	
221 clusters	14,408 unique source values	37,346 rows	

We review the town names after the automatic standardisation, and we manually modify the ones which require further work, relying on the clustering options.

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies Full Data

Clustering options Search values... (/)

Row count	Source value	New value
1	Austi	Austi
1	Azzate	Azzate
1	Este	Este
1	Ucieda	Ucieda
1	Yazd	Yazd
1	Yeste	Yeste
1	Yuyao·City	Yuyao·City

10 values · 21 rows

Source value	New value
12 Lisboa	Lisboa
1 012·Lisboa	Lisboa
1 061·Lisboa	Lisboa
1 095·Lisboa	Lisboa
1 1350211·Lisboa	Lisboa
1 1449041·Lisboa	Lisboa
1 293·Lisboa	Lisboa
1 La Zubia	La Zubia
1 Lisboa·1150282	Lisboa
1 Lisboa·1150·282	Lisboa

10 values · 13 rows

Source value	New value
3 Biella	Biella
2 Balle	Balle

1,482 clusters 14,408 unique source values 37,346 rows 9 selected (20 rows)

Standardize

New value: **Lisboa** Revert to source Apply

Source value: Multiple values Row count: 20

Summary

Source column: Town Unique new values: 14269 Source values updated: 139 / 14408 (0.96%) Rows updated: 180 / 37346 (0.48%)

Cancel Add to Recipe

Once we are pleased with the town names, we “Add to Recipe”.

VANILLA FLOW > companies Full Data

Clustering options Search values... (/)

Row count	Source value	New value
48	Tehran	Tehran
42	Torino	Torino
5	Terni	Terni
4	Torreon	Torreon
3	Torun	Torun
2	Dhahran	Dhahran
2	Trani	Trani
1	Duren	Duren
1	Tarnow	Tarnow
1	Tirana	Tirana
1	Tirane	Tirane
1	Torinio	Torinio

12 values · 111 rows

Source value	New value
132 Cairo	Cairo
3 Guaro	Guaro
2 Cary	Cary
1 Carre	Carre
1 Carru	Carru
1 Chiari	Chiari
1 Gera	Gera
1 Geria	Geria
1 ~~~	~~~

11 values · 145 rows

Source value	New value
--------------	-----------

Select a row to edit.

Summary

Source column: Town Unique new values: 14261 Source values updated: 147 / 14408 (1.02%) Rows updated: 188 / 37346 (0.50%)

Cancel Add to Recipe

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

New Step Recipe X

Company ID 3 RBC Company ID 4 RBC Town ZIP Code

20,050 Categories 14,261 Categories 12,026 Categories

Madrid 28760
 Barbalos De Huebra Salamanca 37455
 445339 Hw sung City GyeonggiDo 445-338
 Biebergemund 63599
 Eggenfelden 84307
 EssenbachAltheim 84051
 Vigo 36214
 Elche 3205
 Paterna 46980
 Arnedo 26580
 Zhejiang 325600
 Oropesa Toledo 45560
 ESB46022711 Valencia 46809
 FR23509496998 Saint Herblain 44806
 Madrid 28045
 Paris 75002
 Reventin 38121
 Limoges 87100
 Mauguio 34130
 ESB93010981 Malaga 29006

8 Columns 49,788 Rows 1 Data Type

17 Standardize Town

Our file should contain a row per company. We are going to deduplicate the rows if all fields have the same values.

VANILLA FLOW > companies > Full Data

New Step Recipe X

Company ID 3 RBC Company ID 4 RBC Town ZIP Code

20,050 Categories 14,261 Categories 12,026 Categories

Madrid 28760
 Barbalos De Huebra Salamanca 37455
 445339 Hw sung City GyeonggiDo 445-338
 Biebergemund 63599
 Eggenfelden 84307
 EssenbachAltheim 84051
 Vigo 36214
 Elche 3205
 Paterna 46980
 Arnedo 26580
 Zhejiang 325600
 Oropesa Toledo 45560
 ESB46022711 Valencia 46809
 FR23509496998 Saint Herblain 44806
 Madrid 28045
 Paris 75002
 Reventin 38121
 Limoges 87100
 Mauguio 34130
 ESB93010981 Malaga 29006

8 Columns 49,788 Rows 1 Data Type

17 Standardize Town

Inconsistent company names demo Standardising values in a text column Celia Muriel – celiamuriel.com



Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

RBC	Country	RBC	Company Name	RBC	Company ID 1	RBC
	165 Categories		44,449 Categories		22,191 Categories	
ES	Ingeteam Power Technology S.A.	A95663852				
ES	Ingeteam Power Technology S.A.	A95663852				
ES	Ingeteam Power Technology S.A.	A95663852				
ES	Ingeteam Power Technology S.A.	A95663852				
ES	Ingeteam Power Technology S.A.	A95663852				
CL	Ingeteam SpA	76198448-9				
AU	Ingeteam Australia Pty. Ltd.	51166870168				
PA	Ingeteam Panamá SA	2619428-1-836502				
RO	Ingeteam Service S.R.L.					
ES	QUANTUM RENEWABLE ENERGY SL	B71219323				
MX	Tai Durango Dos, S.A.P.I. de C.V	TDD1204178Y8				
MX	Tai Durango Cuatro, S.A.P.I. de C.V	TDC120417R14				
PH	Ingeteam Philippines Inc.	8996479000				
ES	BIZKAIA BUSINESS CAPITAL SL	B95810552				
ES	BIZKAIA BUSINESS CAPITAL-1 SL EN CONSTITUCIÓN	B95813069				
ES	BIZKAIA BUSINESS CAPITAL-2 SL EN CONSTITUCIÓN	B95826236				
UY	Ingeteam Uruguay S.A.	21-762707-0016				
ES	GLOBAL INCITATUS SLU	B87251633				
GB	Ingeteam UK Limited	10192806				
ES	INGETEAM R&D EUROPE SL	B95859336				

8 Columns 49,386 Rows 1 Data Type

18 Remove duplicate rows

After all the transformations and cleaning we did on the data, we have 49,386 rows to upload and 44,449 different company names. There are duplicate company names because they have different IDs or they are placed on different cities.

VANILLA FLOW > companies > Full Data

RBC	Country	RBC	Company Name	RBC	Company ID 1	RBC
	165 Categories		44,449 Categories		22,191 Categories	
ES	Ingeteam Power Technology S.A.	A95663852				
ES	Ingeteam Power Technology S.A.	A95663852				
ES	Ingeteam Power Technology S.A.	A95663852				
ES	Ingeteam Power Technology S.A.	A95663852				
ES	Ingeteam Power Technology S.A.	A95663852				
CL	Ingeteam SpA	76198448-9				
AU	Ingeteam Australia Pty. Ltd.	51166870168				
PA	Ingeteam Panamá SA	2619428-1-836502				
RO	Ingeteam Service S.R.L.					
ES	QUANTUM RENEWABLE ENERGY SL	B71219323				
MX	Tai Durango Dos, S.A.P.I. de C.V	TDD1204178Y8				
MX	Tai Durango Cuatro, S.A.P.I. de C.V	TDC120417R14				
PH	Ingeteam Philippines Inc.	8996479000				
ES	BIZKAIA BUSINESS CAPITAL SL	B95810552				
ES	BIZKAIA BUSINESS CAPITAL-1 SL EN CONSTITUCIÓN	B95813069				
ES	BIZKAIA BUSINESS CAPITAL-2 SL EN CONSTITUCIÓN	B95826236				
UY	Ingeteam Uruguay S.A.	21-762707-0016				
ES	GLOBAL INCITATUS SLU	B87251633				
GB	Ingeteam UK Limited	10192806				
ES	INGETEAM R&D EUROPE SL	B95859336				

8 Columns 49,386 Rows 1 Data Type

18 Remove duplicate rows

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



We click on “Run Job”.

The screenshot shows a data processing interface with a table containing 49,386 rows and 8 columns. The columns are labeled: RBC, Country, Company Name, RBC, Company ID 1, RBC, Company ID 2, and RBC, Company ID 3. The Company Name column contains numerous entries such as 'Ingeteam Power Technology S.A.', 'Ingeteam Australia Pty. Ltd.', and 'INGETEAM R&D EUROPE SL'. The Company ID columns contain numerical values like 'A95663852', 'B71219323', and 'B95859336'. On the left side, there are various icons for filtering, sorting, and saving data. In the top right corner, there is a 'Run Job' button with a red arrow pointing to it. The bottom of the interface shows statistics: 165 Categories, 44,449 Categories, 22,191 Categories, 28 Categories, and 33 Categories.

We click on “Create-CSV”. We are going to change this publishing action to append to our BigQuery table.

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



Run Job on Dataflow

Options

Profile results

When enabled, this will generate a profile of your results

Publishing Actions

Add Publishing Action

Actions	Location	Settings
Create-CSV	gs://dataprep-staging-fb698ded-225e-430c-83e8-db35094f7017/celia@heladoscuro.com/jobrun/companies.csv	no compression, multiple files, with quotes, with delimiter: ,

Dataflow Execution Settings

Region: europe-west1

Zone: Auto Zone

Machine Type: n1-standard-1

Advanced Settings ▾

Cancel Run Job

We select BigQuery

Publishing Action

Choose a file or folder

GCS

BigQuery

Output Directory: dataprep-staging-fb698ded-225e-430c-83e8-db35094f7017/celia@heladoscuro.com/jobrun

Create new file Parameterize destination: companies

Output Format: CSV

More options ▾

Cancel Update

We click on our BigQuery dataset.

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



Publishing Action

Choose a table

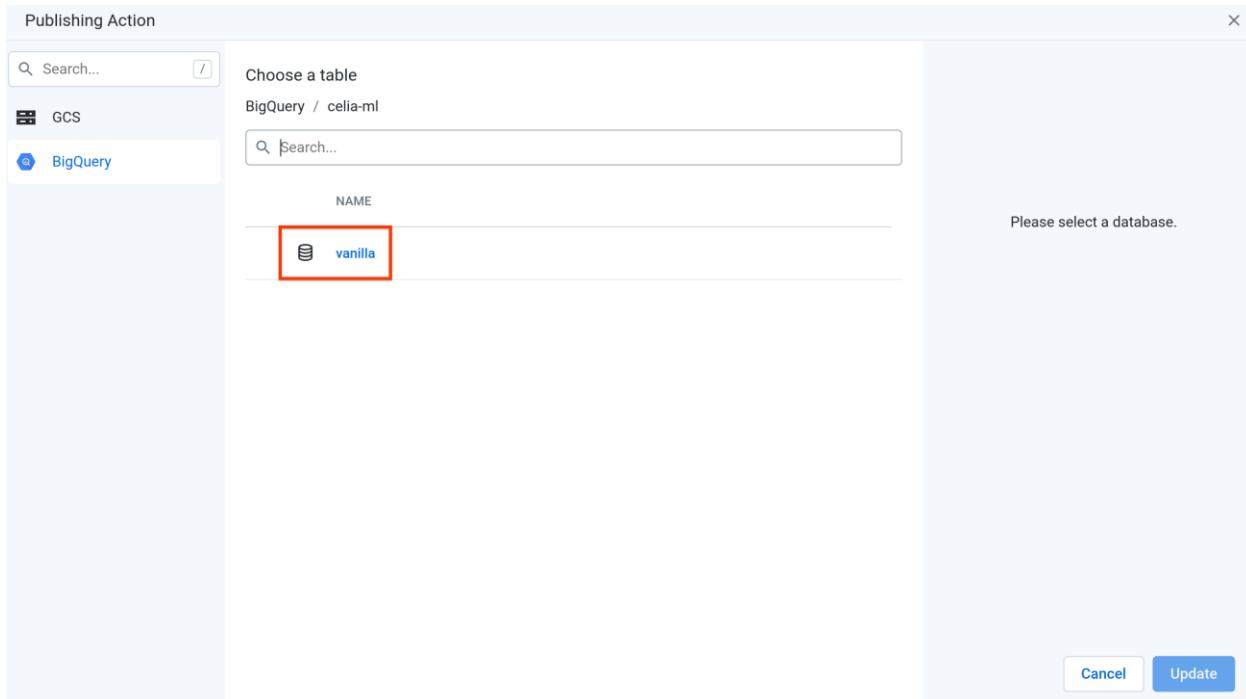
BigQuery / celia-ml

NAME

vanilla

Please select a database.

Cancel Update



We try to append the data in our CSV file to our BigQuery Table. But we can't click on "Update" because we have an error message which says that the fields in the CSV file can't have spaces. So we cancel the publishing action and we come back to our recipe to edit the column names.

Publishing Action

Choose a table

BigQuery / celia-ml / vanilla

NAME SIZE LAST UPDATED

✓ companies 8 Columns 0 Rows

When publishing to BigQuery output(s), column names must begin with a letter or an underscore and can only contain letters, underscores, and digits. Invalid column names: 'Company Name', 'Company ID 1', 'Company ID 2', 'Company ID 3', 'Company ID 4', 'ZIP Code' in companies.

Existing table Back

companies

Output Database

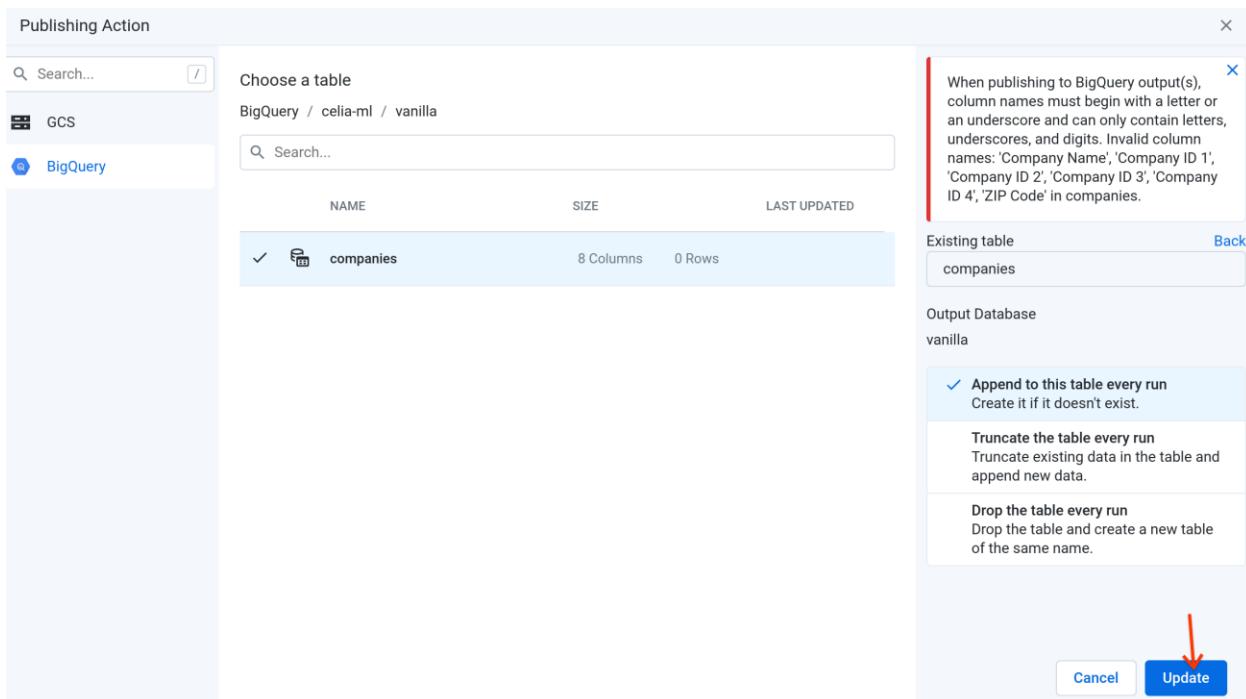
vanilla

✓ Append to this table every run
Create it if it doesn't exist.

Truncate the table every run
Truncate existing data in the table and append new data.

Drop the table every run
Drop the table and create a new table of the same name.

Cancel Update



Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



Publishing Action

Choose a table

BigQuery / celia-ml / vanilla

Existing table companies

Output Database vanilla

Append to this table every run

Truncate the table every run

Drop the table every run

Cancel Update

When publishing to BigQuery output(s), column names must begin with a letter or an underscore and can only contain letters, underscores, and digits. Invalid column names: 'Company Name', 'Company ID 1', 'Company ID 2', 'Company ID 3', 'Company ID 4', 'ZIP Code' in companies.

Run Job on Dataflow

Options

Profile results

When enabled, this will generate a profile of your results

Publishing Actions

Add Publishing Action

Actions	Location	Settings
Create-CSV	gs://dataprep-staging-fb698ded-225e-430c-83e8-db35094f7017/cella@heladoscuro.com/jobrun/companies.csv	no compression, multiple files, with quotes, with delimiter: ,

Dataflow Execution Settings

Region

Zone

Machine Type

Advanced Settings

Cancel Run Job

We rename the CSV column names.

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies < Full Data

Run Job

RBC Country RBC Company Name RBC Company ID 1 RBC

165 Categories 44,449 Categories Ingeteam Power Technology S.A.

ES Ingeteam SpA

CL Ingeteam Australia Pty. Ltd.

AU Ingeteam Panamá SA

PA Ingeteam Service S.R.L.

RO QUANTUM-RENEWABLE ENERGY-SL

ES Tai-Durango Dos, S. A. P. I. de C.

MX Tai-Durango Cuatro, S.A.P.I. de C.V

PH Ingeteam Philippines Inc.

ES BIZKAIA BUSINESS CAPITAL-SL

ES BIZKAIA BUSINESS CAPITAL-1 SL EN CONSTITUCIÓN

UY BIZKAIA BUSINESS CAPITAL-2 SL EN CONSTITUCIÓN

ES Ingeteam Uruguay S.A.

ES GLOBAL INCITATUS SLU

GB Ingeteam UK Limited

ES INGETEAM R&D EUROPE SL

INGETEAM R&D EUROPE SL

8 Columns 49,386 Rows 1 Data Type

Rename

- Change type >
- Move >
- Hide
- Format >
- Calculate >
- Create column from examples
- Group by >
- Pivot
- Restructure >
- Filter rows >
- Replace >
- Standardize...
- Extract >
- Split column >
- Column Details
- Show related Steps in Recipe

VANILLA FLOW > companies < Full Data

Run Job

RBC Country RBC company_name RBC Company ID 1 RBC

165 Categories 44,449 Categories 22,191 Categories 28 Categories

ES Ingeteam Power Technology S.A. A95663852

ES Ingeteam SpA 76198448-9

CL Ingeteam Australia Pty. Ltd. 51166870168

AU Ingeteam Panamá SA 2619428-1-836502

PA Ingeteam Service S.R.L. 21219323

RO QUANTUM-RENEWABLE ENERGY-SL B71204178Y8

ES Tai-Durango Dos, S. A. P. I. de C. TDD120417R14

MX Tai-Durango Cuatro, S.A.P.I. de C.V TDC120417R14

PH Ingeteam Philippines Inc. 8996479000

ES BIZKAIA BUSINESS CAPITAL-SL B95810552

ES BIZKAIA BUSINESS CAPITAL-1 SL EN CONSTITUCIÓN B95813069

ES BIZKAIA BUSINESS CAPITAL-2 SL EN CONSTITUCIÓN B95826236

UY Ingeteam Uruguay S.A. 21-762707-0016

ES GLOBAL INCITATUS SLU B87251633

GB Ingeteam UK Limited 10192806

ES INGETEAM R&D EUROPE SL B95859336

INGETEAM R&D EUROPE SL B95859336

8 Columns 49,386 Rows 1 Data Type

Preview

Rename columns

Option required

Manual rename

Specify the new name for each column

Columns (1) Add

ABC Company Name company_name

Cancel Add

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies > Full Data

8 Columns 49,386 Rows 1 Data Type

company_id3	RBC	company_id4	RBC	town	RBC	zipcode
es	20,050 Categories		14,261 Categories	Albacete	2006	
	ESA95663852			Sesma	31293	
	ESA95663852			Sarriguren	31621	
	ESA95663852			Zamudio	48170	
	ESA95663852			Minano Mayor	1510	
				Las Condes	7550000	
				North Wollongong	2500	
				Distrito De Panama	20335	
				Bucuresti Sector 2	31160	
				Orkoien	11550	
				Ciudad De Mexico	11550	
				Ciudad De Mexico	11550	
				Makati City	1200	
				Zamudio	48170	
				Zamudio	48170	
				Montevideo	11200	
				Madrid	28002	
				Woking	GU21 6LQ	
				Zamudio	48170	

New Step Recipe X

- 16 Remove symbols from Town
- 17 Standardize Town
- 18 Remove duplicate rows
- 19 Rename Company Name to 'company_name' →
- 20 Rename Country to 'country'
- 21 Rename Company ID 1 to 'Company_id1'
- 22 Rename Company_id1 to 'company_id1'
- 23 Rename Company ID 2 to 'company_id2'
- 24 Rename Company ID 3 to 'company_id3'
- 25 Rename Company ID 4 to 'company_id4'
- 26 Rename Town to 'town'
- 27 Rename ZIP Code to 'zipcode'

Once all the CSV columns have the right name format, we click on “Run Job” again.

VANILLA FLOW > companies > Full Data

8 Columns 49,386 Rows 1 Data Type

company_id3	RBC	company_id4	RBC	town	RBC	zipcode
es	20,050 Categories		14,261 Categories	Albacete	2006	
	ESA95663852			Sesma	31293	
	ESA95663852			Sarriguren	31621	
	ESA95663852			Zamudio	48170	
	ESA95663852			Minano Mayor	1510	
				Las Condes	7550000	
				North Wollongong	2500	
				Distrito De Panama	20335	
				Bucuresti Sector 2	31160	
				Orkoien	11550	
				Ciudad De Mexico	11550	
				Ciudad De Mexico	11550	
				Makati City	1200	
				Zamudio	48170	
				Zamudio	48170	
				Montevideo	11200	
				Madrid	28002	
				Woking	GU21 6LQ	
				Zamudio	48170	

New Step Recipe X

- 16 Remove symbols from Town
- 17 Standardize Town
- 18 Remove duplicate rows
- 19 Rename Company Name to 'company_name'
- 20 Rename Country to 'country'
- 21 Rename Company ID 1 to 'Company_id1'
- 22 Rename Company_id1 to 'company_id1'
- 23 Rename Company ID 2 to 'company_id2'
- 24 Rename Company ID 3 to 'company_id3'
- 25 Rename Company ID 4 to 'company_id4'
- 26 Rename Town to 'town'
- 27 Rename ZIP Code to 'zipcode'

↓ Run Job

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



Run Job on Dataflow

Options

Profile results

When enabled, this will generate a profile of your results

Publishing Actions

Actions	Location	Settings
Create-CSV	gs://dataprep-staging-fb698ded-225e-430c-83e8-db35094f7017/celia@heladoscuro.com/jobrun/companies.csv	no compression, multiple files, with quotes, with delimiter: ,

Dataflow Execution Settings

Region

europe-west1

Zone

Auto Zone

Machine Type

n1-standard-1

Advanced Settings ▾

Cancel Run Job

Publishing Action

Choose a file or folder

GCS

BigQuery

Search... / celia@heladoscuro.com /jobrun

Create Folder

Search...

NAME	SIZE	LAST UPDATED
.data_prep_temp_c5f86652-b...		
20200601_Companies_1.csv/		
p_20200601_Companies_570...		
p_20200601_Companies_570...		
p_20200601_Companies_570...		

Create a new file Parameterize destination

companies

Output Directory

dataprep-staging-fb698ded-225e-430c-83e8-db35094f7017/celia@heladoscuro.com/jobrun

Data Storage Format

CSV

Create new file every run
 Create a new file with an incremental number appended to the name (e.g. companies_2.csv)

Append to this file every run
 Create it if it doesn't exist.

Replace this file every run
 Create it if it doesn't exist.

More options ▾

Cancel Update

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



Publishing Action

Choose a table

BigQuery / celia-ml

Please select a database.

NAME

 vanilla

Cancel Update

Publishing Action

Choose a table

BigQuery / celia-ml / vanilla

Create a new table

or choose an existing table

NAME SIZE LAST UPDATED

 companies	8 Columns	0 Rows
---	-----------	--------

Cancel Update

We select “append to this table” and we click on “Update”.

Inconsistent company names demo

Standardising values in a text column

Celia Muriel – celiamuriel.com



Publishing Action

Choose a table

BigQuery / celia-ml / vanilla

Existing table

Output Database

vanilla

Click here to show columns that don't match

Append to this table every run

Create it if it doesn't exist.

Truncate the table every run

Truncate existing data in the table and append new data.

Drop the table every run

Drop the table and create a new table

Cancel Update

A red arrow points from the "Drop the table every run" section down to the "Update" button.

Click on “Run Job”.

Run Job on Dataflow

Options

Profile results

When enabled, this will generate a profile of your results

Publishing Actions

Add Publishing Action

Actions	Location	Settings
Append-BigQuer	celia-ml:vanilla.companies	Create table if it does not exist; Append

Dataflow Execution Settings

Region

europe-west1

Zone

Auto Zone

Machine Type

n1-standard-1

Advanced Settings

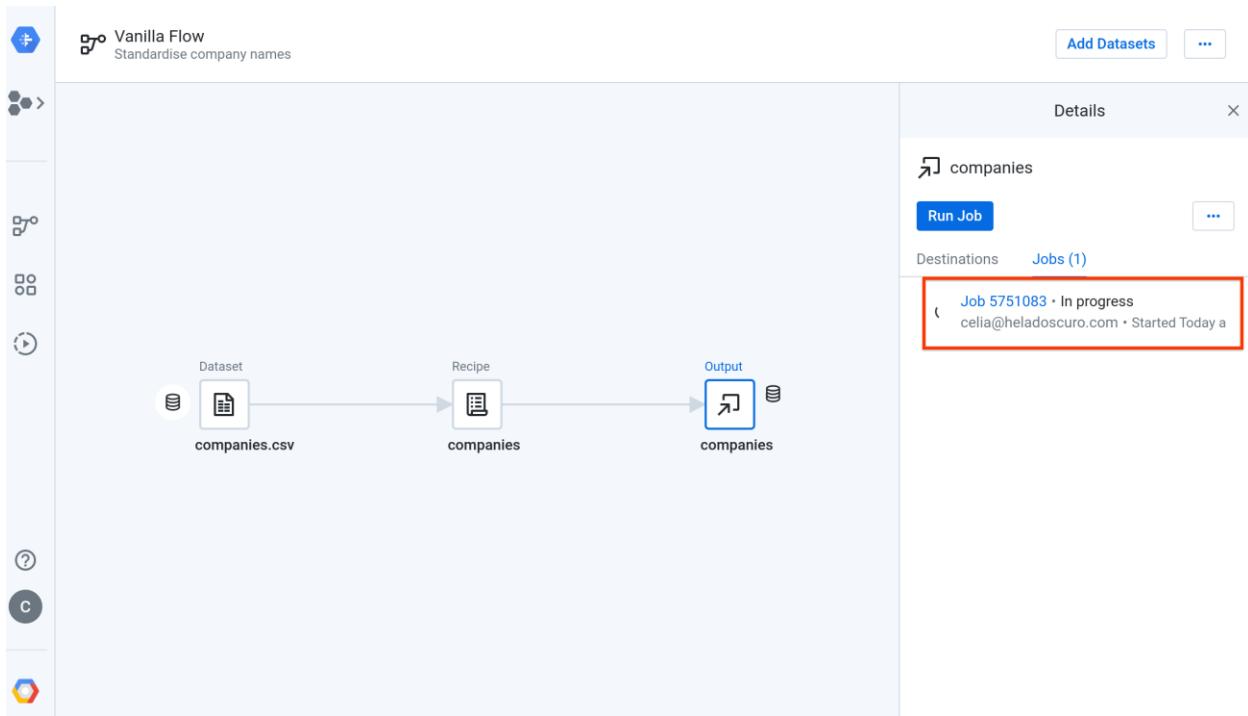
Cancel Run Job

A red arrow points from the "Run Job" button down to the "Run Job" button in the Dataflow interface.

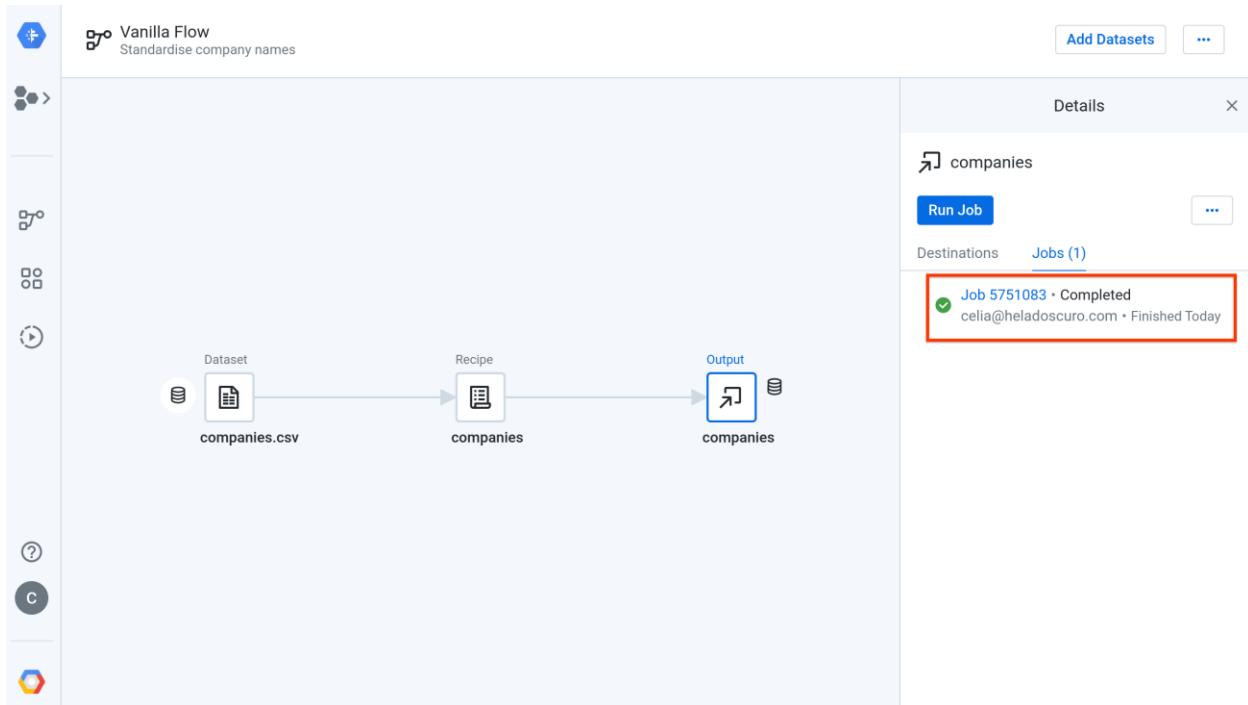
Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



Cloud Dataflow is the combination of Trifecta software for data preparation and Cloud Dataflow to upload the data. When we hit on “Run Job”, we trigger a Cloud Dataflow job which actually copies the data from the CSV file to the BigQuery table.



Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



Once the job completes successfully, we check in BigQuery that we upload the standardised companies data into our table.

The screenshot shows the Google Cloud Platform BigQuery interface. The left sidebar lists datasets: 'celia-ml' (expanded, showing 'vanilla' and 'companies'), 'bigquery-public-data', and 'bigquery-public-data'. The 'companies' dataset is selected. The main area is the 'Query editor' where the following SQL query is run:

```
1 select *
2 from `vanilla.companies`;
```

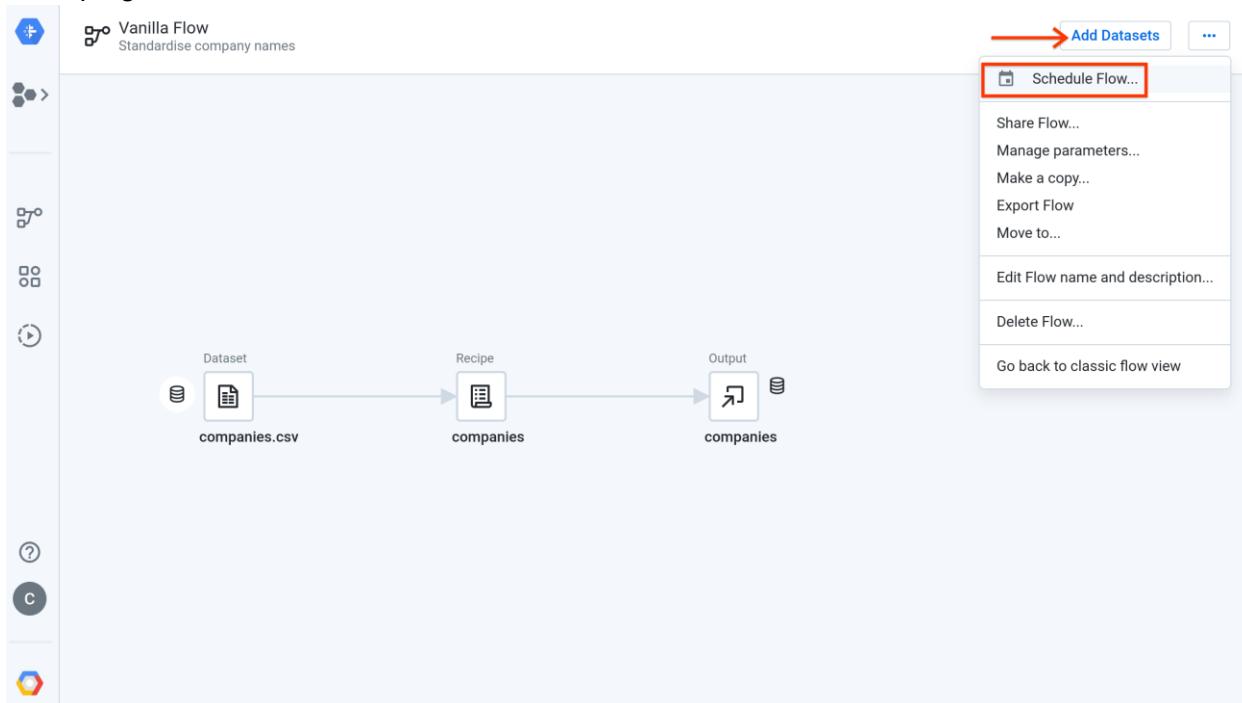
The 'Query results' section shows the output of the query. The table has columns: Row, company_name, company_id1, company_id2, company_id3, company_id4, and country. The data includes:

Row	company_name	company_id1	company_id2	company_id3	company_id4	country
1	(N	CARGO SOLAR POWER PRIVATE LIMITED				AAECC0306N
2	AD	FORCES ELECTRIQUES D'ANDORRA				
3	AE	ORANGE OVERSEAS FZE				
4	AE	Haris Al Afaq Ltd.				
5	AE	KALHOUR OILFIELD EQUIPMENTS LTD				
6	AE	Amplex Emirates LLC				
7	AE	ECHO CARGO & SHIPPING LLC				
8	AE	STERLING & WILSON ME SOLAR ENE				
9	AE	Mahindra Susten Private Limited - Dubai Branch				

At the bottom, it says 'Rows per page: 100 1 - 100 of 49386 First page < > Last page'.



If we want to schedule our job to run regularly to upload the data from our CSV file to our BigQuery Table, we come back to the Dataprep flow. We open it, and we click on the 3 dots on the top right corner. We click on “Schedule Flow”.



Recipe and flow

The complete [recipe](#) and [flow](#) with all transformations done to prepare the companies.csv data to load are available in these links.

Other option

If we have a master table in BigQuery with the standardised company names, we can also create a flow in Cloud Dataprep where we join the source CSV file and the master table. Then we can compare records using the [DOUBLEMETAPHONEEQUALS](#) function.

The [double metaphone algorithm](#) is based on [phonetic coding](#). It encodes an English word phonetically by reducing them to a combination of 12 consonant sounds. It also attempts to encode non-English words. It returns two codes if a word has two plausible pronunciations.

If we want to add a function as the DOUBLEMETAPHONEEQUALS, we must open the recipe again, and proceed as shown in the screenshots below.

Inconsistent company names demo
 Standardising values in a text column
 Celia Muriel – celiamuriel.com



VANILLA FLOW > companies < Full Data

New Step ↑ Recipe X

1 Keep rows where ISMISMATCHED((ZIP Code), [Integer])
 2 Standardize Company Name
 3 Trim quotes from Company Name
 4 Trim whitespace from Company Name

RBC	Country	RBC	Company Name	RBC	Company ID 1	RBC
	133 Categories		INDUSTRIAS BLOOM LYRA S/A ASTROLABIO HENRYKA		471 Categories	No
BR			Faculdade Leão Sampaio Fernando			
BR			CPFL Renováveis			
PL			SIEMENS-GAMESA RENEWABLE ENERGY POLAND SP. ZOO			
PE			RED ELÉCTRICA ANDINA S.A.C. REA		20511261571	
BR			Alstom Energias Renováveis Ltda			
BR			Siemens Gamesa Energia Renovavel LT		69119386000151	
GB			NRS Group UK Ltd Noel Regan and Sons Building & Green Energy Solutions			
HT			ACN Articulos y Construcciones Eléctric		J0210000184742	
NI			ACCIONA WINDPOWER BRASIL			
BR			ACCIONA WINDPOWER BRASIL-COM. IND. EXP			
BR			ALEMINTSA, S.A.		991297480001	
EC			Solarclarity BV Att.: Derek Durham		NL820757743B01	
NL			General Electric Canada		869542407RT0001	
CA			FLORIDA OIL & GAS TECHNOLOGIES INC Isabel Sousa			
US			SIA-MARINE SYSTEMS			
LV			Canadian Solar Solutions Inc.			
CA			British Solar Renewables Ltd.		159976146	
GB			NV Texels Eigen Stoomboot Ondernemi			
NL			Arturo Enrique Solano Urrutia-TECNO			
SV						

8 Columns 4,633 Rows 2 Data Types

VANILLA FLOW > companies < Full Data

Search Transformations X

New Step ↑ Recipe X

Search... (Ctrl+k)

Formulas

- Scale to min max
- Scale a column to a specific min max range
- One hot encode
- Create a column for each unique value indicating i...
- Scale to mean
- Scale a column to zero mean and unit variance
- Bin column
- Bin values into ranges of equal or custom size
- New formula
- Create a new column from the result of a formula
- Select
- Derive a new table with an arbitrary schema from ...
- Edit with formula
- Set one or more columns to the result of a formula
- Window
- Perform calculations across multiple ordered rows
- Schema

RBC	Country	RBC	Company Name	RBC	Company ID 1	RBC
	133 Categories		INDUSTRIAS BLOOM LYRA S/A ASTROLABIO HENRYKA		471 Categories	No
BR			Faculdade Leão Sampaio Fernando			
BR			CPFL Renováveis			
PL			SIEMENS-GAMESA RENEWABLE ENERGY POLAND SP. ZOO			
PE			RED ELÉCTRICA ANDINA S.A.C. REA		20511261571	
BR			Alstom Energias Renováveis Ltda			
BR			Siemens Gamesa Energia Renovavel LT		69119386000151	
GB			NRS Group UK Ltd Noel Regan and Sons Building & Green Energy Solutions			
HT			ACN Articulos y Construcciones Eléctric		J0210000184742	
NI			ACCIONA WINDPOWER BRASIL			
BR			ACCIONA WINDPOWER BRASIL-COM. IND. EXP			
BR			ALEMINTSA, S.A.		991297480001	
EC			Solarclarity BV Att.: Derek Durham		NL820757743B01	
NL			General Electric Canada		869542407RT0001	
CA			FLORIDA OIL & GAS TECHNOLOGIES INC Isabel Sousa			
US			SIA-MARINE SYSTEMS			
LV			Canadian Solar Solutions Inc.			
CA			British Solar Renewables Ltd.		159976146	
GB			NV Texels Eigen Stoomboot Ondernemi			
NL			Arturo Enrique Solano Urrutia-TECNO			
SV						

8 Columns 4,633 Rows 2 Data Types

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



VANILLA FLOW >
companies ▾
Full Data

Run Job

Search Transformations

do

DOUBLEMETAPHONE
(Function) Returns a single array containing the phonetic representation of each string in the input array.

DOUBLEMETAPHONEEQUALS
(Function) Checks if two strings match phonetically.

DOMAIN
(Function) Returns the domain from a valid URL.

Filter not equals
Filter rows which do not equal a value

Comment
Add a comment to your recipe

FLOOR
(Function) Rounds the value down to the nearest integer.

ISMISMATCHED
(Function) Checks if a value does not conform to a regular expression.

Replace missing
Replace cells with missing values with a new value

IFMISMATCHED
(Function) Returns a supplied value if the input value does not match a regular expression.

RBC	Country	RBC	Company Name	RBC	Company ID 1	RBC
133 Categories	INDUSTRIAL DESIGN LTD. OF ASIA LTD INDUSTRY	4508 Categories	Faculdade Leão Sampaio Fernando	471 Categories	No	
BR	CPFL Renováveis	SIEMENS GAMESA RENEWABLE ENERGY POLAND SP. ZOO	RED·ELÉCTRICA ANDINA S.A.C.·REA	20511261571		
PL	Alstom Energias Renováveis Ltda	Siemens Gamesa Energia Renovável LT		69119386000151		
PE	NRS Group UK Ltd Noel Regan and Sons Building & Green Energy Solutions	Green Energy Solutions				
BR	ACN Articulos y Construcciones Eléctric	J0210000184742				
BR	ACCIONA WINDPOWER BRASIL					
BR	ACCIONA WINDPOWER BRASIL COM. IND. EXP					
EC	ALEMINDA, S.A.	991297480001				
NL	Solarclarlity BV Att.: Derek Durham	NL820757743B01				
CA	General Electric Canada	869542407RT0001				
US	FLORIDA OIL & GAS TECHNOLOGIES INC Isabel Sousa					
LV	SIA MARINE SYSTEMS					
CA	Canadian Solar Solutions Inc.					
GB	British Solar Renewables Ltd.	159976146				
NL	NV Texels Eigen Stoomboot Onderneming					
SV	Arturo Enrique Solano Urrutia-TECNO					

VANILLA FLOW > companies > Full Data

New formula

Formula type required

Single row formula

Create a new column from a single row formula

Formula required

`DOUBLEMETAPHONEQUALS()`

`DOUBLEMETAPHONEQUALS(string1, string2, match_threshold)`

Checks if two strings match phonetically using the Double Metaphone phonetic encoding algorithm. [Learn more](#)

Example

`DOUBLEMETAPHONEQUALS(Smith, Schmidt, normal)`

string1 string

The first string you want to compare. This can be a string, a function returning a string, or a column containing strings.

Browse

Columns Functions Metadata

RBC	Country	RBC	Company Name	RBC	Company ID 1	RBC
133 Categories	BR	4,508 Categories	Faculdade Leão Sampaio Fernando	471 Categories	No	
	BR		CPFL Renováveis			
	PL		SIEMENS GAMESA-RENEWABLE ENERGY POLAND SP. ZOO			
	PE		RED ELÉCTRICA ANDINA S.A.C. REA	20511261571		
	BR		Aistom Energias Renováveis Ltda			
	BR		Siemens Gamesa Energia Renovavel LT	69119386000151		
	GB		NRS Group-UK Ltd Noel Regan and Sons Building & Green Energy Solutions			
	HT		ACN Articulos y Construcciones Eléctric	J0210000184742		
	NI		ACCIONA WINDPOWER BRASIL			
	BR		ACCIONA WINDPOWER BRASIL COM. IND. EXP			
	EC		ALEMINSA, S.A.	991297480001		
	NL		Solarclarity BV Att.: Derek Durham	NL820757743B01		
	CA		General Electric Canada	869542407RT0001		
	US		FLORIDA OIL & GAS TECHNOLOGIES INC Isabel Sousa SIA MARINE SYSTEMS			
	LV		Canadian Solar Solutions Inc.			
	CA		British Solar Renewables-Ltd.	159976146		
	GB		NV Texels Eigen Stoomboot Ondernemi			
	NL		Arturo Enrique Solano Urrutia-TECNO			

8 Columns 4,633 Rows 2 Data Types

Other fuzzy matching techniques, such as the [Levenshtein Distance](#), must be implemented by the user as functions in Trifecta to use them.

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



References

Trifacta

[Advanced Data Cleanup Techniques using Cloud Dataprep](#). [Cloud Next '19](#).
[Google Cloud Platform](#). Accessed November 7th, 2021.

[Quickstart](#). [Cloud Dataprep by Trifacta](#). [Google Cloud](#). Accessed November 7th, 2021.

[Enabling and Disabling Dataprep by Trifacta](#). [Cloud Dataprep by Trifacta](#). [Google Cloud](#). Accessed November 7th, 2021.

[User Profile Page](#). [Documentation for Trifacta Wrangler](#). Accessed November 7th, 2021.

[Overview of Standardization](#). [Documentation for Trifacta Wrangler](#). Accessed November 7th, 2021.

[DOUBLEMETAPHONE Function](#). [Documentation for Trifacta Wrangler](#). Accessed November 7th, 2021.

[DOUBLEMETAPHONEEQUALS Function](#). [Documentation for Trifacta Wrangler](#). Accessed November 7th, 2021.

[Wrangle Language](#). [Documentation for Trifacta Wrangler](#). Accessed November 7th, 2021.

[Compare Strings](#). [Documentation for Trifacta Wrangler](#). Accessed November 7th, 2021.

[Join Window](#). [Documentation for Trifacta Wrangler](#). Accessed November 7th, 2021.

David McNamara. [December '19 Wrangler Release – Rapid Target Fuzzy Match, UI Improvements, Downloadable Profiles](#). [Trifacta Blog](#). December 18th, 2019. Accessed November 7th, 2021.

Bertrand Cariou. [New AI-driven features in Dataprep enhance the wrangling experience](#). [Google Cloud Blog](#). April 8th, 2020. Accessed November 7th, 2021.

Other

Celia Muriel. [Fuzzy Matching or approximate string matching](#). Available on December 13th, 2021.

Inconsistent company names demo
Standardising values in a text column
Celia Muriel – celiamuriel.com



[Using the bq command-line tool](#). [BigQuery](#). [Google Cloud](#). Accessed November 7th, 2021.

[Dataflow](#). [Google Cloud](#). Accessed November 7th, 2021.