

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	44
Other option	44
References	47
Trifacta	47
Other	47

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.

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

CELIA MURIEL

The screenshot shows the DataWrangler interface. On the left is a sidebar with icons for Flows, Datasets, Projects, Help, and Configuration. The main area is titled "Flows" with buttons for "Create..." and "...". It shows tabs for "All Flows", "Owned by me", and "Shared with me". A search bar is at the top right. In the center, there's a large icon of a flowchart with the text "Create a flow to wrangle your data." and a "Create Flow" button with a red arrow pointing to it. A modal window titled "Create Flow" is open in the foreground. It has fields for "Flow Name" (containing "Vanilla Flow") and "Flow Description" (containing "Standardise company names"). At the bottom are "Cancel" and "Create" buttons, with a red arrow pointing to the "Create" button.

Create Flow

Flow Name

Vanilla Flow

Flow Description

Standardise company names

Cancel Create

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



The screenshot shows the Dataflow interface. On the left is a sidebar with various icons: a hexagon with a plus sign, a group of people, a gear, a question mark, a circle with a 'C', and a circular arrow. The main area displays a flow titled "Vanilla Flow" with the subtitle "Standardise company names". A large icon of a document with a grid pattern is centered. Below it, the text "Import data before wrangling in this Flow." is displayed. To the right of this text is a button labeled "Import & Add Datasets" with a red arrow pointing to it. In the top right corner of the main area, there is a three-dot menu icon.

The screenshot shows the "Import Data and Add to Flow" dialog box. On the left, there is a sidebar with "Search..." and "Upload" buttons, and sections for "GCS" (selected), "BigQuery", and "BigTable". The main area has a search bar and a "Choose a file or folder" section. Under "GCS", there is a "Create Dataset with Parameters" button and a search bar. To the right, a message says "0 New Datasets" and "Choose data to import." Below this, a table lists datasets: "dataprep-staging-fb...", "dataprep-staging-fb...", and "vanilla_bucket" (which is highlighted with a red box). At the bottom right are "Import & Add to Flow" and "Cancel" buttons.



Inconsistent Company Names Demo

CELIA MURIEL

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

CELIA MURIEL

The screenshot shows the Vanilla Flow interface. On the left, there's a sidebar with various icons. In the center, a dataset named "companies.csv" is displayed under the "Dataset" section. To the right, a "Details" panel is open for "companies.csv". The panel includes a preview of the CSV file, which contains several rows of data. A red arrow points to the "Add" button in the preview header. Below the preview, there are details about the dataset such as Type (GCS), Location (gs://vanilla_bucket/companies.csv), and File Size (2.79MB).

This screenshot is similar to the one above, showing the Vanilla Flow interface with a dataset named "companies.csv". However, a modal window titled "Recipe" is overlaid on the "Details" panel. The "Recipe" modal has three options: "Join" and "Union" (which is currently selected) under the "Review" tab. A red arrow points to the "Union" option. The rest of the interface, including the dataset preview and detailed information, remains the same.



Inconsistent Company Names Demo

CELIA MURIEL

The screenshot shows the Dataprep interface with a sidebar containing various icons. The main area displays a 'Vanilla Flow' titled 'Standardise company names'. A flow diagram shows a 'Dataset' node ('companies.csv') connected to a 'Recipe' node ('companies'). The 'Recipe' node has an 'Edit Recipe' button highlighted with a red arrow. The 'Details' panel on the right shows the recipe name 'companies', a note about no steps, and metadata like 'Steps: 0', 'Updated: Today at 8:40 PM', and 'Created: Today at 8:40 PM'.

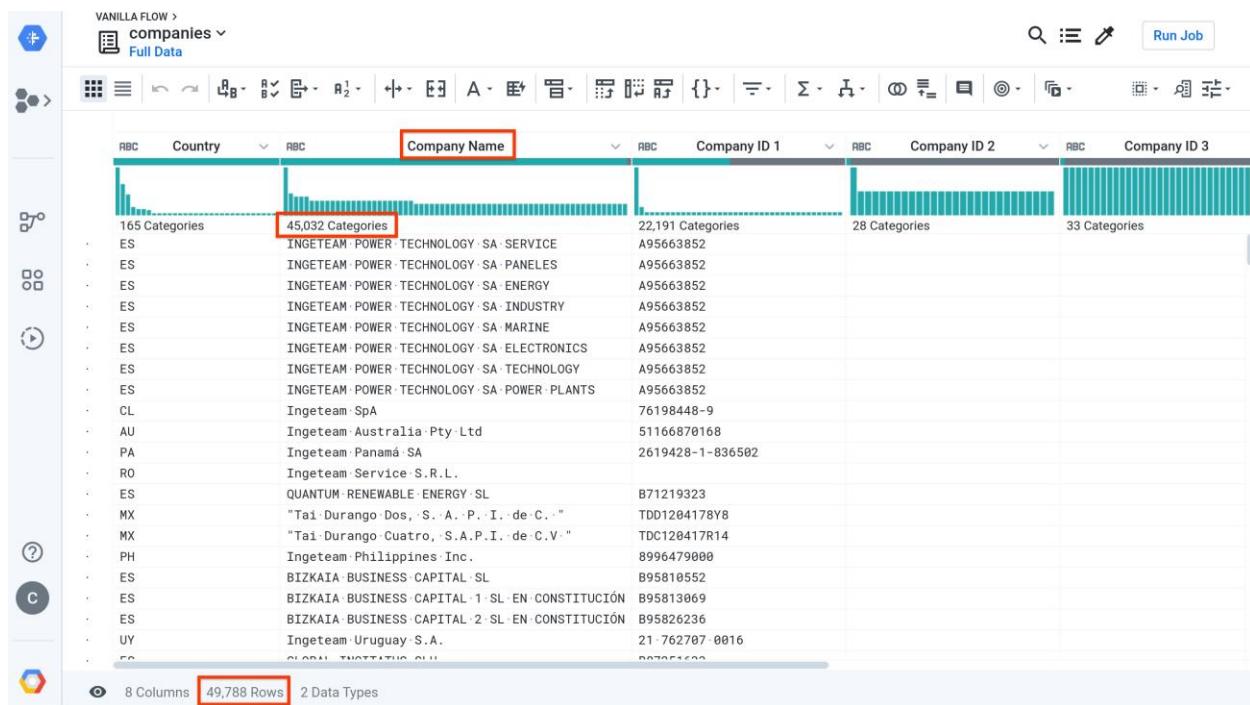
When we edit the recipe for the first time, it shows us the data in the CSV file. Dataprep allows to quickly analyse the data and add the transformations easily.

The screenshot shows the Dataprep Data Preview screen for the 'companies' dataset. The preview table has columns: RBC, Country, RBC, Company Name, RBC, Company ID 1, RBC, Company ID 2, RBC, Company ID 3. The preview includes histograms for each column and a list of rows. 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.



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

CELIA MURIEL

This screenshot shows a data processing interface with a context menu open over a column named "ZIP Code". The menu is titled "# ZIP Code" and includes options like "Rename", "Change type", "Move", "Format", "Calculate", "Create column from examples", "Group by", "Pivot", "Restructure", "Filter rows", "Replace", "Standardize...", "Extract", "Split column", "Column Details", and "Show related Steps in Recipe". The "Change type" option is highlighted with a red box. The "ZIP Code" column itself contains various values such as "es", "20,050 Categories", "16,469 Categories", "Albacete", "Sesma", "Sarriguren", "Zamudio", "Miñano Mayor", "Las Condes", "North Wollongong", "Distrito de Panamá", "Bucuresti, Sector 2", "Orkoién", "CIUDAD DE MÉXICO", "Makati City", "Zamudio", "Montevideo", and "RO34091550".

This screenshot shows a data processing interface with a context menu open over a column named "ZIP Code". The menu is identical to the one in the previous screenshot, with the "Change type" option highlighted. The "ZIP Code" column contains the same set of values as the first screenshot.

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 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.



Inconsistent Company Names Demo

CELIA MURIEL

VANILLA FLOW > companies > Full Data

Run Job

RBC	Country	RBC	Company Name	RBC	Company ID 1	RBC	Company ID 2	RBC	Company ID 3
165 Categories	45,032 Categories	22,191 Categories	28 Categories	33 Categories					
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. 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 INSTITUTE SRL	B97051600							

8 Columns 49,788 Rows 1 Data Type

We search for “no usar” (do not use).

VANILLA FLOW > companies > Full Data

Run Job

Columns Rows

no usar

Clear all filters

RBC	Country	RBC	Company Name	RBC	Comp
165 Categories	45,032 Categories	22,191 Categories			
ES	"NO USAR" - ACERINOX, S.A. "	A28250777			
MX	(NO USAR) OCEAN EUROPE SA DE CV				
PA	Solarcentury (NO USAR)	155586449			
ES	ETECNIC SCP (NO USAR)	B55527824			
RU	no usar				
MA	NO USAR-ACWA POWER BOUDOUR	2069851B			
ES	Montajes y Mantenimientos 2020 SL NO USAR	B91810747			
ES	NO USAR SISTECA INNOVATION SL	B31926108			
ES	"CROVI, S.A. (NO USAR)"	A08403545			
ES	Noratel Spain S.L. (NO USAR)	B29869146			
ES	"Urilfiltr, S.L. NO USAR NO USAR"	B20589321			
ES	"NO-USAR(BZZ MOBILIARIO,S.L.)"	B20557880			
ES	"Releco, S.A. ***NO USAR***VER DISAILECO 206949"	A28672624			
GB	Southco EURpe Ltd.(NO USAR)	396121055			
ES	NO USAR * EPCS ELECTRONIC COMP. SUSTITUIR POR 208102	A79252151			
IT	Falco Electronics (No usar)	108209990			
ES	"NO USAR" PROSEGUR CIA. SEGURIDAD, S. "	A28430882			
NL	NO USAR CADENCE SOLUTIONS PROVIDER	NL812545771B01			
GB	"Southco Manufacturing, Ltd.(NO USAR"	GB785399463			
US	NO USAR * KEMET Electronics Corp				
ES	"NO USAR" CO. ADOBYKO AG				

8 Columns 50 Rows 1 Data Type

We highlight one of the “no usar” 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”.



Inconsistent Company Names Demo

CELIA MURIEL

VANILLA FLOW > companies > Full Data

Source to be dropped Preview

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

9 Columns 50 Rows 1 Data Type

Show only affected Columns Rows

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)+'

VANILLA FLOW > companies > Full Data

New Step Recipe

Change ZIP Code type to String

Replace matches of 'NO USAR' from Company Name with "

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

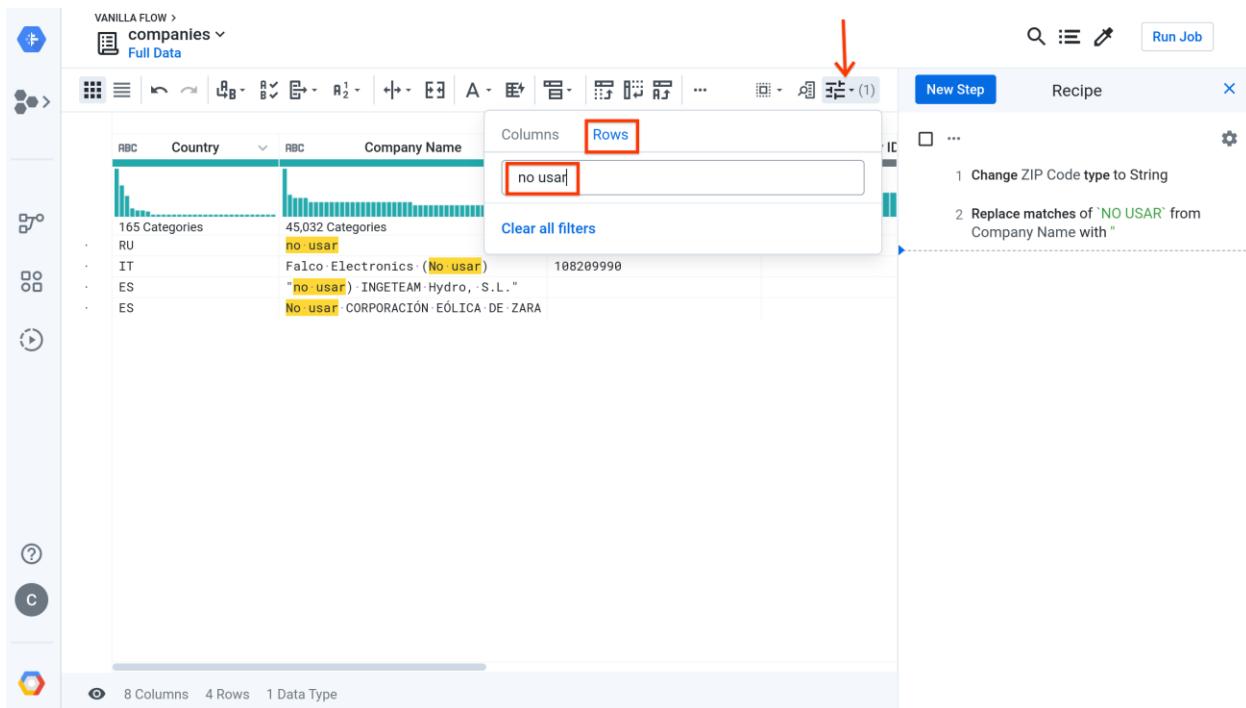
8 Columns 4 Rows 1 Data Type

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



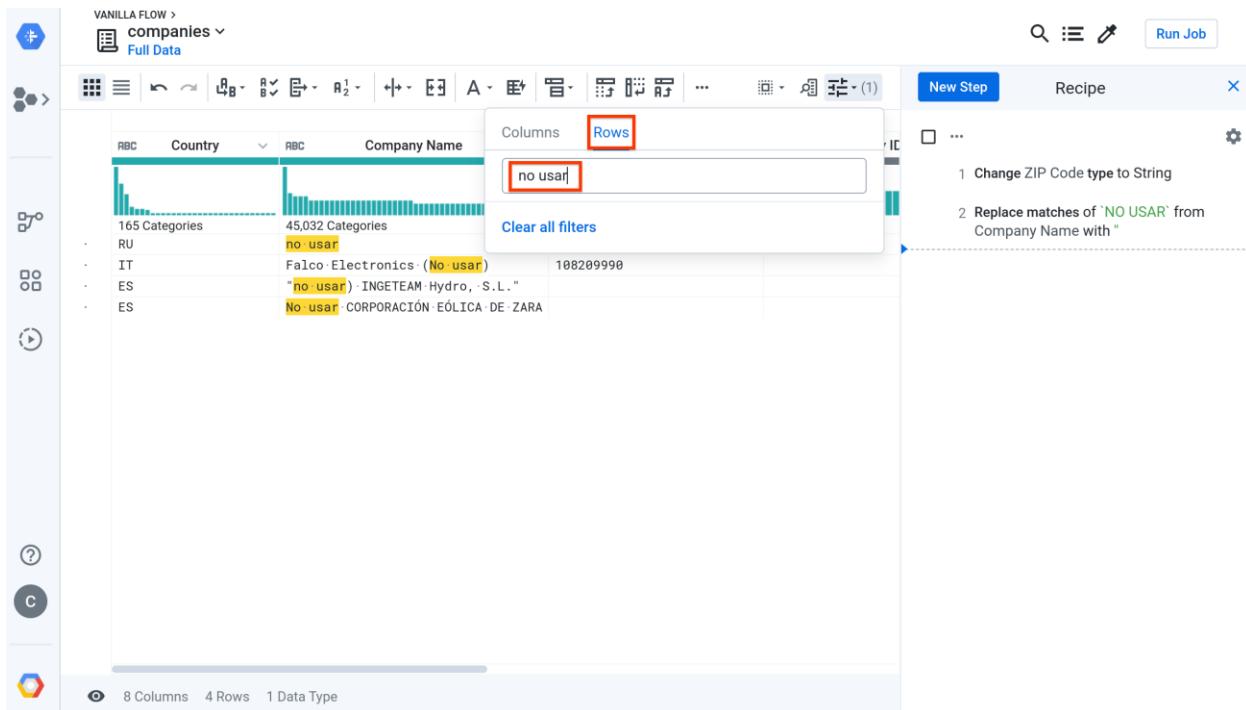
Inconsistent Company Names Demo

CELIA MURIEL



The screenshot shows the Cloud Dataflow interface with a search bar containing 'no usar'. A red box highlights the 'Rows' tab in the search dropdown. Below the search bar, a table displays company data with several entries containing the string 'no usar' highlighted in yellow. The interface includes a sidebar with various icons and a recipe editor on the right.

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.



This screenshot is identical to the one above, showing the Cloud Dataflow interface with a search bar containing 'no usar'. A red box highlights the 'Rows' tab in the search dropdown. The table below shows company data with 'no usar' entries highlighted in yellow. The sidebar and recipe editor are also visible.



Inconsistent Company Names Demo

CELIA MURIEL

The screenshot shows a data processing interface with a sidebar containing icons for file operations, a search bar, and a 'Run Job' button. The main area displays a table with two columns: 'Country' and 'Company Name'. The 'Country' column has three rows: 'RBC', 'IT', and 'ES'. The 'Company Name' column has two rows: 'Falco Electronics (No user)' and 'CORPORACIÓN EÓLICA DE ZARA'. A red box highlights the 'Rows' tab in the top right corner of the table view. A search bar at the bottom of the table contains the text 'No user'. To the right of the table, a 'Recipe' panel lists three steps:

- 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 ''

At the bottom of the interface, it says '8 Columns 2 Rows 1 Data Type'.

This screenshot is similar to the one above, showing the same data processing interface. The table now has three rows in the 'Company Name' column: 'MICROBERRI', 'Microsoft Ireland', and 'Bergen Engines AS'. A red box highlights the 'Rows' tab in the top right corner of the table view. A search bar at the bottom of the table contains the text 'no utiliz'. To the right of the table, a 'Recipe' panel lists four steps:

- 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 ''

At the bottom of the interface, it says '8 Columns 3 Rows 1 Data Type'.



Inconsistent Company Names Demo

CELIA MURIEL

The screenshot shows a data processing interface with a sidebar containing icons for file operations, a search bar, and a 'Run Job' button. The main area displays a data grid with columns: RBC, Country, and Company Name. A red box highlights the 'Rows' tab in the top navigation bar. In the Company Name column, several entries are highlighted in yellow, including 'non usare', 'non_usare*****MERKUR Ueberseehandel GmbH', and 'non_usare'. A red box also highlights the 'non usare' entry in the Company Name column. Below the grid, a message says 'Clear all filters'. The bottom status bar indicates 8 Columns, 2 Rows, and 1 Data Type.

The screenshot shows a data processing interface with a sidebar containing icons for file operations, a search bar, and a 'Run Job' button. The main area displays a data grid with columns: RBC, Country, and Company Name. A red box highlights the 'Rows' tab in the top navigation bar. In the Company Name column, several entries are highlighted in yellow, including 'non utilizzare', 'GLOBAL ENERGY S.R.L. - DA N', 'NON UTILIZZARE - CIEL IMPIANTI S.R.L.', 'TOTAL SOLAR - NON UTILIZZARE', '****non utilizzare***JSC Hydroenergy', and '*** non utilizzareCAMERA DI COMMERCIO UFFICIALE SPAGNOLA III >'. A red box also highlights the 'non utilizzare' entry in the Company Name column. Below the grid, a message says 'Clear all filters'. The bottom status bar indicates 8 Columns, 5 Rows, and 1 Data Type.

Now we standardise the company names.



Inconsistent Company Names Demo

CELIA MURIEL

VANILLA FLOW > companies > Full Data

Details

RBC Company Name

Quality

- Valid 49781 99.99%
- Mismatched 0 0%
- Missing 7 0.01%

Unique Values

Value	Count
NULL	13
Koop-Brinkmann-GmbH	6
PERSONA-FISICA	5
GRUPPO SONNEDIX	5

Patterns

Pattern	Count
{upper}+ {upper}+ {upper}+	6,216
{upper}+ {upper}+{delim}{upper}+{delim}	4,826
{upper}+ {upper}+	3,941
{upper}+	1,237
{upper}+ {upper}+ {upper}.{upper}.{upper}	1,075

Show pattern details...

Suggestions

Delete columns

8 Columns 49,788 Rows 1 Data Type

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.

VANILLA FLOW > companies > Full Data

Clustering options

Search values... (/)

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.

6 values · 9 rows

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.

4 values · 4 rows

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

4 values · 4 rows

Source value	New value
4 INGETEAM GmbH	INGETEAM GmbH
2 INGETEAM GMBH	INGETEAM GMBH

4 values · 8 rows

Select a row to edit.

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

1,010 clusters 45,031 unique source values 49,788 rows



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

Row count	Source value	New value	
<input checked="" type="checkbox"/>	2 Ingeteam S.R.L.	Ingeteam S.R.L.	6 values · 9 rows
<input checked="" type="checkbox"/>	2 INGETEAM-s.r.l.	Ingeteam S.R.L.	
<input checked="" type="checkbox"/>	2 INGETEAM-S.R.L.	Ingeteam S.R.L.	
<input checked="" type="checkbox"/>	1 "INGETEAM, S.R.L."	Ingeteam S.R.L.	
<input checked="" type="checkbox"/>	1 ***INGETEAM S.R.L.	Ingeteam S.R.L.	
<input checked="" type="checkbox"/>	1 Ingeteam s.r.l.	Ingeteam S.R.L.	
<input type="checkbox"/>	1 "GENERAL-ELECTRIC INTERNATIONAL, Inc."	"GENERAL-ELECTRIC INTERNATIONAL, Inc"	4 values · 4 rows
<input type="checkbox"/>	1 "GENERAL-ELECTRIC INTERNATIONAL, Inc"	"GENERAL-ELECTRIC INTERNATIONAL, Inc"	
<input type="checkbox"/>	1 GENERAL-ELECTRIC INTERNATIONAL INC	GENERAL-ELECTRIC INTERNATIONAL INC	
<input type="checkbox"/>	1 GENERAL-ELECTRIC INTERNATIONAL INC.	GENERAL-ELECTRIC INTERNATIONAL INC.	
<input type="checkbox"/>	1 Gamesa Wind GmbH	Gamesa Wind GmbH	4 values · 4 rows
<input type="checkbox"/>	1 Gamesa Wind GMBH	Gamesa Wind GMBH	
<input type="checkbox"/>	1 GAMESA WIND GmbH	GAMESA WIND GmbH	
<input type="checkbox"/>	1 GAMESA WIND GMBH	GAMESA WIND GMBH	
<input type="checkbox"/>	4 INGETEAM GmbH	INGETEAM GmbH	4 values · 8 rows
<input type="checkbox"/>	2 INGETEAM GMBH	INGETEAM GMBH	
1,010 clusters	45,031 unique source values	49,781 rows	6 selected (9 rows)

New value: Ingeteam S.R.L.

Standardize
 Revert to source
 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%)

We confirm we were clustering on similar strings.

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

Row count	Source value	New value	
<input checked="" type="checkbox"/>	2 Ingeteam S.R.L.	Ingeteam S.R.L.	6 values · 9 rows
<input checked="" type="checkbox"/>	2 INGETEAM-s.r.l.	Ingeteam S.R.L.	
<input checked="" type="checkbox"/>	2 INGETEAM-S.R.L.	Ingeteam S.R.L.	
<input checked="" type="checkbox"/>	1 "INGETEAM, S.R.L."	Ingeteam S.R.L.	
<input checked="" type="checkbox"/>	1 ***INGETEAM S.R.L.	Ingeteam S.R.L.	
<input checked="" type="checkbox"/>	1 Ingeteam s.r.l.	Ingeteam S.R.L.	
<input type="checkbox"/>	1 "GENERAL-ELECTRIC INTERNATIONAL, Inc."	"GENERAL-ELECTRIC INTERNATIONAL, Inc"	4 values · 4 rows
<input type="checkbox"/>	1 "GENERAL-ELECTRIC INTERNATIONAL, Inc"	"GENERAL-ELECTRIC INTERNATIONAL, Inc"	
<input type="checkbox"/>	1 GENERAL-ELECTRIC INTERNATIONAL INC	GENERAL ELECTRIC INTERNATIONAL INC	
<input type="checkbox"/>	1 GENERAL-ELECTRIC INTERNATIONAL INC.	GENERAL ELECTRIC INTERNATIONAL INC.	
<input type="checkbox"/>	1 Gamesa Wind GmbH	Gamesa Wind GmbH	4 values · 4 rows
<input type="checkbox"/>	1 Gamesa Wind GMBH	Gamesa Wind GMBH	
<input type="checkbox"/>	1 GAMESA WIND GmbH	GAMESA WIND GmbH	
<input type="checkbox"/>	1 GAMESA WIND GMBH	GAMESA WIND GMBH	
<input type="checkbox"/>	4 INGETEAM GmbH	INGETEAM GmbH	4 values · 8 rows
<input type="checkbox"/>	2 INGETEAM GMBH	INGETEAM GMBH	
1,010 clusters	45,031 unique source values	49,781 rows	6 selected (9 rows)



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.

VANILLA FLOW > companies > Full Data

Row count > Source value New value

	Source value	New value
<input type="checkbox"/>	1 YC2-ENERJY	YC2-ENERJY
<input checked="" type="checkbox"/>	2 Ingeteam S.R.L.	Ingeteam S.R.L.
<input checked="" type="checkbox"/>	2 INGETEAM s.r.l.	Ingeteam S.R.L.
<input checked="" type="checkbox"/>	2 INGETEAM S.R.L.	Ingeteam S.R.L.
<input type="checkbox"/>	2 INGETEAM SRL	INGETEAM SRL
<input checked="" type="checkbox"/>	1 "INGETEAM, S.R.L.."	Ingeteam S.R.L.
<input checked="" type="checkbox"/>	1 ***INGETEAM S.R.L.	Ingeteam S.R.L.
<input checked="" type="checkbox"/>	1 Ingeteam s.r.l.	Ingeteam S.R.L.

7 values · 11 rows

	Source value	New value
<input type="checkbox"/>	2 "ESF-SPANIEN-0301, S.L."	"ESF-SPANIEN-0301, S.L."
<input type="checkbox"/>	2 "ESF-SPANIEN-0303, S.L."	"ESF-SPANIEN-0303, S.L."
<input type="checkbox"/>	2 ESF-Spanien 0302 S.L.U.	ESF-Spanien 0302 S.L.U.
<input type="checkbox"/>	2 ESF-Spanien 0306 S.L.U.	ESF-Spanien 0306 S.L.U.
<input type="checkbox"/>	2 ESF-Spanien 0308 S.L.U.	ESF-Spanien 0308 S.L.U.
<input type="checkbox"/>	2 ESF-Spanien 0311 S.L.U.	ESF-Spanien 0311 S.L.U.
<input type="checkbox"/>	1 ESF-SPANIEN 0424 SL	ESF-SPANIEN 0424 SL

7 values · 13 rows

	Source value	New value
<input type="checkbox"/>	2 IS ENERGY SRL	IS ENERGY SRL
<input checked="" type="checkbox"/>	1 "EOSA-ENERGIA, S.R.L.."	"EOSA-ENERGIA, S.R.L.."
<input checked="" type="checkbox"/>	1 IS ENERGY SRL	IS ENERGY SRL

7 values · 8 rows

2,410 clusters 45,031 unique source values 49,781 rows 6 selected (9 rows)

VANILLA FLOW > companies > Full Data

Row count > Source value New value

	Source value	New value
<input type="checkbox"/>	1 YC2-ENERJY	YC2-ENERJY
<input checked="" type="checkbox"/>	2 Ingeteam S.R.L.	Ingeteam S.R.L.
<input checked="" type="checkbox"/>	2 INGETEAM s.r.l.	Ingeteam S.R.L.
<input checked="" type="checkbox"/>	2 INGETEAM S.R.L.	Ingeteam S.R.L.
<input checked="" type="checkbox"/>	2 INGETEAM SRL	Ingeteam S.R.L.
<input checked="" type="checkbox"/>	1 "INGETEAM, S.R.L.."	Ingeteam S.R.L.
<input checked="" type="checkbox"/>	1 ***INGETEAM S.R.L.	Ingeteam S.R.L.
<input checked="" type="checkbox"/>	1 Ingeteam s.r.l.	Ingeteam S.R.L.

7 values · 11 rows

	Source value	New value
<input type="checkbox"/>	2 "ESF-SPANIEN-0301, S.L."	"ESF-SPANIEN-0301, S.L."
<input type="checkbox"/>	2 "ESF-SPANIEN-0303, S.L."	"ESF-SPANIEN-0303, S.L."
<input type="checkbox"/>	2 ESF-Spanien 0302 S.L.U.	ESF-Spanien 0302 S.L.U.
<input type="checkbox"/>	2 ESF-Spanien 0306 S.L.U.	ESF-Spanien 0306 S.L.U.
<input type="checkbox"/>	2 ESF-Spanien 0308 S.L.U.	ESF-Spanien 0308 S.L.U.
<input type="checkbox"/>	2 ESF-Spanien 0311 S.L.U.	ESF-Spanien 0311 S.L.U.
<input type="checkbox"/>	1 ESF-SPANIEN 0424 SL	ESF-SPANIEN 0424 SL

7 values · 13 rows

	Source value	New value
<input type="checkbox"/>	2 IS ENERGY SRL	IS ENERGY SRL
<input checked="" type="checkbox"/>	1 "EOSA-ENERGIA, S.R.L.."	"EOSA-ENERGIA, S.R.L.."
<input checked="" type="checkbox"/>	1 IS ENERGY SRL	IS ENERGY SRL

7 values · 8 rows

2,410 clusters 45,031 unique source values 49,781 rows 7 selected (11 rows)



Inconsistent Company Names Demo

CELIA MURIEL

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 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.

The screenshot shows the Dataedo Standardize interface for the 'companies' dataset. In the 'Source value' column, several rows for 'Ingeteam S.R.L.' are selected and highlighted with a red box. The 'New value' column shows the corrected value 'Ingeteam S.R.L.'. On the right, the 'Standardize' panel shows the input 'Ingeteam S.R.L.' and the 'Apply' button, which is highlighted with a blue arrow. The summary at the bottom indicates 7 selected (11 rows) and 11 rows updated.

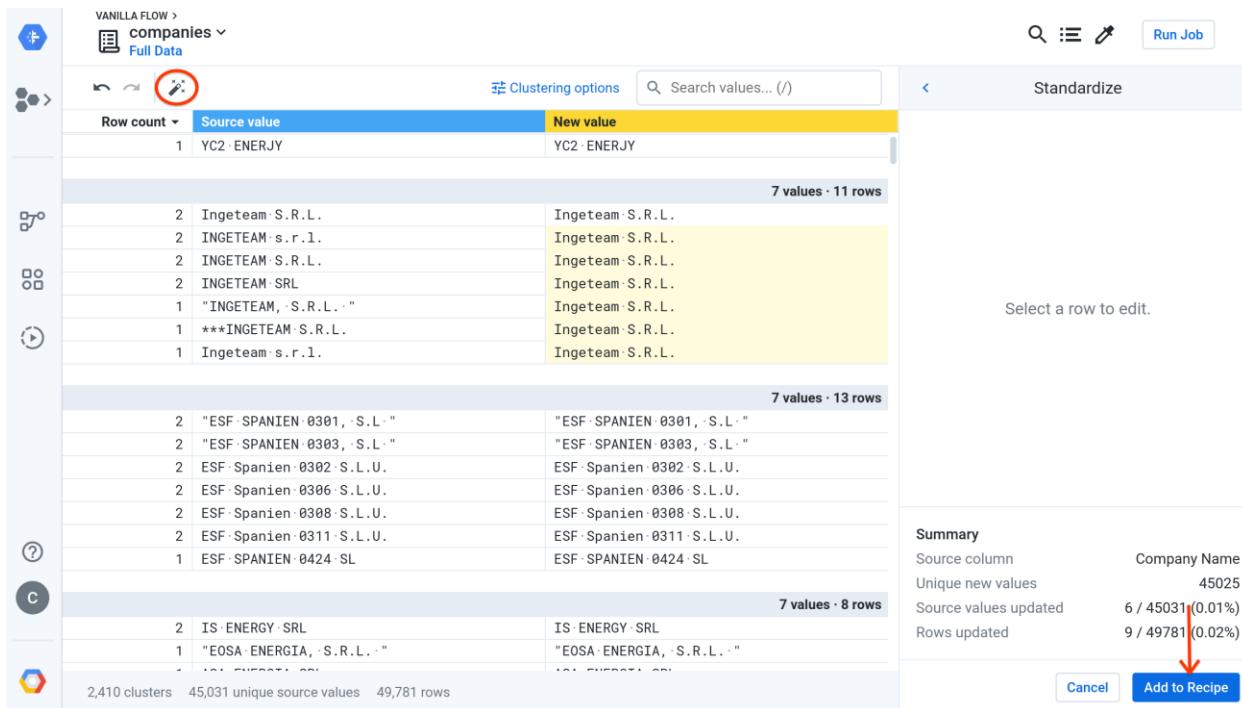
This screenshot shows the same standardization process after the 'Ingeteam S.R.L.' row has been successfully applied. The 'Summary' section now shows 0 / 49781 (0.00%) rows updated. A note 'Select a row to edit.' is displayed above the summary. The rest of the interface is identical to the previous screenshot.



Inconsistent Company Names Demo

CELIA MURIEL

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.



The screenshot shows the Cloud Dataprep interface for the 'companies' flow, specifically the 'Full Data' step. A red circle highlights the 'Standardize' button in the top right corner of the main table area. The table has two columns: 'Source value' and 'New value'. The 'Source value' column contains various company names like 'YC2-ENERJY', 'Ingeteam S.R.L.', and 'ESF-SPANIEN'. The 'New value' column shows the standardized versions. Below the table, a summary section provides statistics: 2,410 clusters, 45,031 unique source values, and 49,781 rows. On the right, a sidebar displays a summary of changes made:

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

At the bottom right, there are 'Cancel' and 'Add to Recipe' buttons, with a red arrow pointing to the 'Add to Recipe' button.

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



Inconsistent Company Names Demo

CELIA MURIEL

VANILLA FLOW > companies Full Data

Clustering options Search values... (/)

Standardize

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.

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

New Step Recipe X

RBC Country RBC Company Name RBC Company ID 1 RBC

165 Categories 44,686 Categories 22,191 Categories 28 Categories

RBC	Country	RBC	Company Name	RBC	Company ID 1	RBC
ES	SANJOSE TECNOLOGIAS SA	A36409918				
ES	"ALCAZAREN SOSTENIBLE, S.L."	B37468170				
KR	"SAMSUNG Heavy Industries Co., LTD."	6128500343				
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"	B97822308				
ES	ENERGIAS ALTERNATIVAS ARNEDO	B26382143				
CN	"HUAYI WIND ENERGY CO., LTD - ZHEJIANG HEWIND	3308736897797				
ES	HERNANDEZ RUIZ ANDRES	70413750				
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				

8 Columns 49,788 Rows 1 Data Type

10 Standardize Company Name

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



Inconsistent Company Names Demo

CELIA MURIEL

VANILLA FLOW > companies > Full Data

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 ENERGÍAS ALTERNATIVAS ARNEDO
- CN "HUAYI WIND ENERGY CO., LTD ZHEJIANG HEWIND
- ES HERNANDEZ RUIZ ANDRES
- ES "DELEGADA TÉCNICA, 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

Context menu for 'Company Name' column:

- 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

Sub-menu for 'Format':

- 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

VANILLA FLOW > companies > Full Data

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 ENERGÍAS ALTERNATIVAS ARNEDO
- CN "HUAYI WIND ENERGY CO., LTD ZHEJIANG HEWIND
- ES HERNANDEZ RUIZ ANDRES
- ES "DELEGADA TÉCNICA, S.L."
- FR Systovi
- ES TERREROS SERRANO ANTONIO
- FR VANITEX - SOLARENT
- FR Ecomotiv
- FR Dja Stock Energie
- FR Solea
- ES ADJ DITEC MALAGA SLL

9 Columns 49,788 Rows 1 Data Type

Text format dialog:

- Columns required
- Multiple
- RBC Company Name
- Format required
- Trim leading and trailing quotes

Note: Remove quotes found at the beginning and end of the text

Add button highlighted with a red arrow.



Inconsistent Company Names Demo

CELIA MURIEL

The screenshot shows a data processing interface with a table of company names and a list of 11 steps to standardize them:

- 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 "
- 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

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

The screenshot shows a data processing interface with a context menu open over a company name column. The 'Format' option is selected, and the 'Trim leading and trailing whitespace' option is highlighted.



Inconsistent Company Names Demo

CELIA MURIEL

VANILLA FLOW > companies > Full Data

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

9 Columns 49,788 Rows 1 Data Type

Show only affected Columns

Text format

Columns required
Multiple

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.	61285080343			
-	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	B97822308			
-	ES	ENERGIAS ALTERNATIVAS ARNEDO	B26382143			
-	CN	HUAYI WIND ENERGY CO., LTD ZHEJIANG HEWIND C	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			

8 Columns 49,788 Rows 1 Data Type

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 useare' 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 UTILIZARE' 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

CELIA MURIEL

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-do, South Korea" Biebergemünd Eggenfelden Essenbach-Altheim VIGO ELCHE Paterna ARNEDO Zhejiang OROPESA (Toledo) Valencia Saint Herblain MADRID Paris Reventin Limoges Mauguiro 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

VANILLA FLOW > companies > Full Data

Source to be dropped Preview

3 RBC Company ID 4 RBC Town RBC Town RBC

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

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

Text format

Columns required
Multiple
RBC Town

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

Cancel Add

9 Columns 49,788 Rows 1 Data Type



Inconsistent Company Names Demo

CELIA MURIEL

VANILLA FLOW >
companies >
Full Data

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

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

Madrid 28760
Barbalos De Huebra (Salamanca) 37455
"445-330 Hwsung City, -Gyeong 445-330
Biebergemünd 63599
Eggenfelden 84307
Essenbach-Altheim 84051
Vigo 36214
Elche 3285
Paterna 46980
Arnedo 26580
Zhejiang 325600
Oropesa (Toledo) 45560
Valencia 46009
Saint Herblain 44806
Madrid 28045
Paris 75002
Reventin 38121
Limoges 87100
Mauguio 34130
Malaga 29006

8 Columns 49,788 Rows 1 Data Type

...
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
13 Rename Country1 to 'Town'
14 Convert text in Town to Propercase

We remove accents.

VANILLA FLOW >
companies >
Full Data

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

20,050 Categories 14,827 Categories

Madrid
Barbalos De Huebra (Salamanca)
"445-330 Hwsung City, -Gyeong
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

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



Inconsistent Company Names Demo

CELIA MURIEL

VANILLA FLOW > companies > Full Data

Company ID 3	RBC	Company ID 4	RBC	Town	RBC	Town
Barbalos	20,050 Categories		14,827 Categories	Madrid	20,050 Categories	Madrid
				Barbalos-De Huebra (Salamanca)		Barbalos-De Huebra (S.
				"445-330 Hwsung City, Gyeong		"445-330 Hwsung City,
				Biebergemund		Biebergemund
				Eggenfelden		Eggenfelden
				Essenbach-Altheim		Essenbach-Altheim
				Vigo		Vigo
				Elche		Elche
				Paterna		Paterna
				Arnedo		Arnedo
				Zhejiang		Zhejiang
				Oropesa (Toledo)		Oropesa (Toledo)
				Valencia		Valencia
				Saint Herblain		Saint Herblain
				Madrid		Madrid
				Paris		Paris
				Reventin		Reventin
				Limoges		Limoges
				Mauguio		Mauguio
				Malaga		Malaga

9 Columns 49,788 Rows 1 Data Type

Show only affected Columns

Text format Run Job

Columns required
Multiple RBC Town

Format required
Remove accents Cancel Add

Remove all accents from the text

VANILLA FLOW > companies > Full Data

Company ID 3	RBC	Company ID 4	RBC	Town	RBC	ZIP Code
Barbalos	20,050 Categories		14,609 Categories	Madrid	28760	
				Barbalos-De Huebra (Salamanca)	37455	
				"445-330 Hwsung City, Gyeong	445-330	
				Biebergemund	63599	
				Eggenfelden	84387	
				Essenbach-Altheim	84051	
				Vigo	36214	
				Elche	3205	
				Paterna	46980	
				Arnedo	26580	
				Zhejiang	325600	
				Oropesa (Toledo)	45560	
				Valencia	46009	
				Saint-Herblain	44886	
				Madrid	28045	
				Paris	75082	
				Reventin	38121	
				Limoges	87180	
				Mauguio	34130	
				Malaga	29006	

8 Columns 49,788 Rows 1 Data Type

New Step Recipe Run Job

... New Step

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 Run Job

We remove symbols.



Inconsistent Company Names Demo

CELIA MURIEL

VANILLA FLOW > companies > Full Data

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

Company ID 3	RBC	Company ID 4	RBC	Town	RBC	ZIP Code	RBC
Companies		20,050 Categories		14,609 Categories			
				Madrid			
				Barbalos De Huebra (Salamanca)			
				"445-330 Hwswing City, Gyeonggi-do"			
				Biebergemund			
				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

Context menu options for the 'Town' column:

- Rename
- Change type >
- Move >
- Hide
- Format** > (highlighted with red box)
- Calculate >
- Create column from examples
- Group by >
- Pivot
- Restructure >
- Filter rows >
- Replace
- Standardize...
- Extract
- Split column >
- Column Details
- Show related Steps in Recipe

Sub-menu for 'Format':

- Convert to UPPERCASE
- Convert to lowercase
- Convert to Proper Case
- Trim leading and trailing whitespace
- Trim leading and trailing quotes
- Remove whitespace
- Remove symbols** (highlighted with red box)
- Remove accents

VANILLA FLOW > companies > Full Data

Source to be dropped Preview

RBC	Company ID 4	RBC	Town	RBC	Town	RBC
			14,609 Categories		14,408 Categories	
			Madrid	Madrid	28	
			Barbalos De Huebra (Salamanca)	Barbalos De Huebra Salamanca	37	
			"445-330 Hwswing City, Gyeonggi-do"	445330 Hwswing City GyeonggiDo	44	
			Biebergemund	Biebergemund	63	
			Eggenfelden	Eggenfelden	84	
			Essenbach-Altheim	EssenbachAltheim	84	
			Vigo	Vigo	36	
			Elche	Elche	32	
			Paterna	Paterna	46	
			Arnedo	Arnedo	26	
			Zhejiang	Zhejiang	32	
			Oropesa (Toledo)	Oropesa Toledo	45	
			Valencia	Valencia	46	
			Saint Herblain	Saint Herblain	44	
			Madrid	Madrid	28	
			Paris	Paris	75	
			Reventin	Reventin	38	
			Limoges	Limoges	87	
			Mauguio	Mauguio	34	
			Malaga	Malaga	29	

9 Columns 49,788 Rows 1 Data Type

Text format dialog:

- Columns: required
- Multiple
- RBC Town** (highlighted with red box)
- Format: required
- Remove symbols (highlighted with red box)

Buttons: Cancel, Add (highlighted with red arrow)



Inconsistent Company Names Demo

CELIA MURIEL

VANILLA FLOW > companies > Full Data

New Step Recipe X

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
- 16 Remove symbols from Town

8 Columns 49,788 Rows 1 Data Type

We standardise the Town field with the Wizard.

VANILLA FLOW > companies > Full Data

8 Columns 49,788 Rows 1 Data Type



Inconsistent Company Names Demo

CELIA MURIEL

VANILLA FLOW > companies > Full Data

Row count: Source value | New value

4 values · 14 rows

3 values · 9 rows

3 values · 6 rows

3 values · 11 rows

3 values · 15 rows

221 clusters 14,408 unique source values 37,346 rows

Select a row to edit.

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

VANILLA FLOW > companies > Full Data

Row count: Source value | New value

4 values · 14 rows

3 values · 9 rows

3 values · 6 rows

3 values · 11 rows

3 values · 15 rows

221 clusters 14,408 unique source values 37,346 rows

Select a row to edit.

Summary

- Source column Town
- Unique new values -
- Source values updated -
- Rows updated -

Auto standardize changed 139 values in 132 clusters. Undo

Cancel Add to Recipe

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

CELIA MURIEL

The screenshot shows a data processing interface for a 'VANILLA FLOW > companies' dataset. The main area displays a table of source values and new values. A specific row for 'Lisboa' is highlighted with a red box, indicating it is being processed. The summary on the right shows 14269 unique new values created from 139 source values.

Source value	New value
Austi	Austi
Azzate	Azzate
Este	Este
Ucieda	Ucieda
Yazd	Yazd
Yeste	Yeste
Yuyao City	Yuyao City
Lisboa	Lisboa
012 Lisboa	Lisboa
061-Lisboa	Lisboa
095 Lisboa	Lisboa
1350211 Lisboa	Lisboa
1449041-Lisboa	Lisboa
293-Lisboa	Lisboa
La Zubia	La Zubia
Lisboa 1150282	Lisboa
Lisboa 1150 282	Lisboa
Biella	Biella
Balle	Balle

Summary:

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

Buttons: Cancel, Add to Recipe

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

The screenshot shows the same data processing interface after the changes have been added to the recipe. The 'Add to Recipe' button is highlighted with a blue arrow. The summary now shows 14261 unique new values created from 147 source values.

Source value	New value
Tehran	Tehran
Torino	Torino
Terni	Terni
Torreon	Torreon
Torun	Torun
Dhahran	Dhahran
Trani	Trani
Duren	Duren
Tarnow	Tarnow
Tirana	Tirana
Tirane	Tirane
Torinio	Torinio
Cairo	Cairo
Guaro	Guaro
Cary	Cary
Carre	Carre
Carru	Carru
Chiari	Chiari
Gera	Gera
Geria	Geria
...	...

Summary:

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

Buttons: Cancel, Add to Recipe



Inconsistent Company Names Demo

CELIA MURIEL

The screenshot shows a data processing interface with a table of company data and a sidebar for a 'Recipe'.

Table 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
				Biebergemund		63599
				Eggenfelden		84307
				EssenbachAltheim		84051
				Vigo		36214
				Elche		3205
				Paterna		46980
				Arnedo		26580
				Zhejiang		325600
				Oropesa Toledo		45560
				Valencia		46009
				Saint Herblain		44806
				Madrid		28045
				Paris		75002
				Reventin		38121
				Limoges		87100
				Mauguio		34130
				Malaga		29006

Recipe Sidebar:

- Replace matches of 'non usare' from Company Name with ''
- Replace matches of 'NON USARE' from Company Name with ''
- Replace matches of 'non utilizzare' from Company Name with ''
- Replace matches of 'NON UTILIZZARE' from Company Name with ''
- Standardize Company Name
- Trim quotes from Company Name
- Trim whitespace from Company Name
- Rename Country1 to 'Town'
- Convert text in Town to Propercase
- Remove accents from Town
- Remove symbols from Town
- Standardize Town

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

The screenshot shows the same data processing interface after deduplicating rows. The table now contains fewer unique rows.

Table Data (Deduplicated):

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
				Biebergemund		63599
				Eggenfelden		84307
				EssenbachAltheim		84051
				Vigo		36214
				Elche		3205
				Paterna		46980
				Arnedo		26580
				Zhejiang		325600
				Oropesa Toledo		45560
				Valencia		46009
				Saint Herblain		44806
				Madrid		28045
				Paris		75002
				Reventin		38121
				Limoges		87100
				Mauguio		34130
				Malaga		29006



Inconsistent Company Names Demo

CELIA MURIEL

VANILLA FLOW > companies > Full Data

Run Job

Search Transformations

deduplicate

Remove duplicate rows

Search documentation for "deduplicate"

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

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

	RBC	Country	RBC	Company Name	RBC	Company ID 1	RBC
165 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				
	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				
	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				
	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				
	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				

8 Columns 49,788 Rows 1 Data Type

VANILLA FLOW > companies > Full Data

Run Job

Remove duplicate rows

Add

Preview

Country Company Name

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

RBC	Country	RBC	Company Name	RBC	Company ID 1	RBC
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				
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				
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				
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				
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				
ES	Ingeteam Power Technology S.A.	A95663852				
ES	Ingeteam Power Technology S.A.	A95663852				
ES	Ingeteam Power Technology S.A.	A95663852				

Show only affected Columns Rows



Inconsistent Company Names Demo

CELIA MURIEL

VANILLA FLOW > companies > Full Data

New Step Recipe X

8 Columns 49,386 Rows 1 Data Type

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

Company Name Company ID 1

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

16 Remove symbols from Town

17 Standardize Town

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

New Step Recipe X

8 Columns 49,386 Rows 1 Data Type

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

Company Name Company ID 1

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

16 Remove symbols from Town

17 Standardize Town

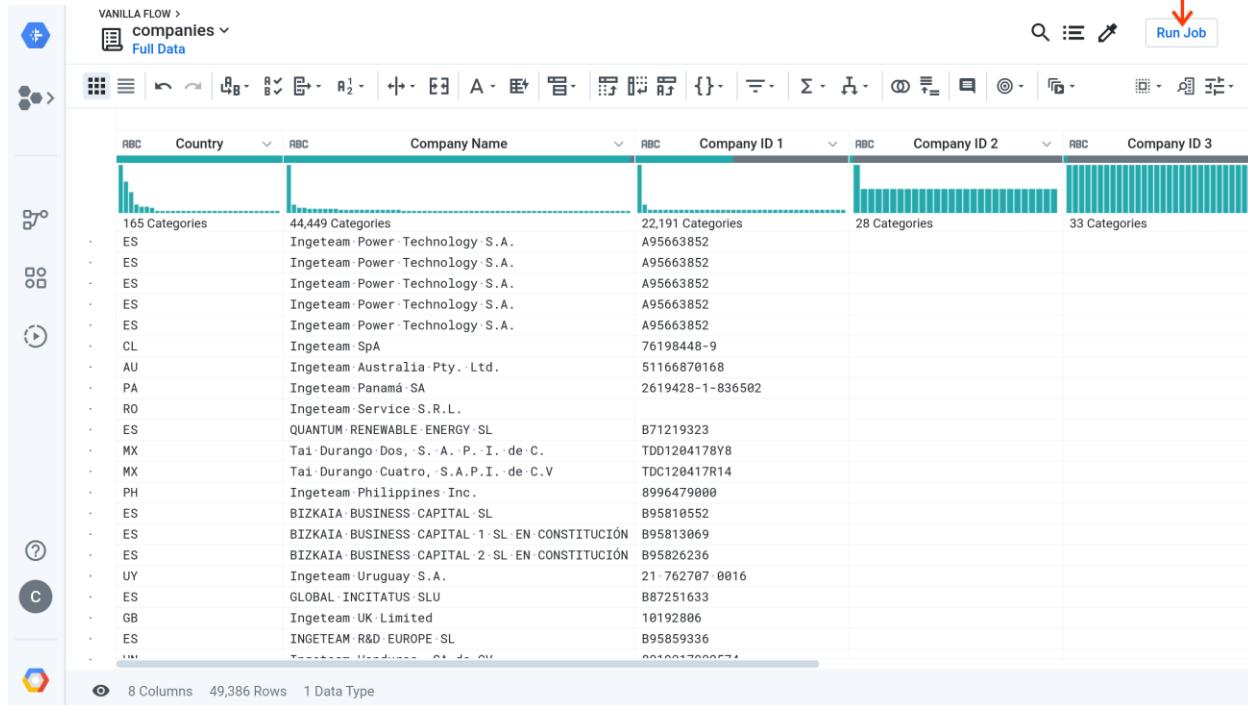
18 Remove duplicate rows

We click on "Run Job".



Inconsistent Company Names Demo

CELIA MURIEL



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

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

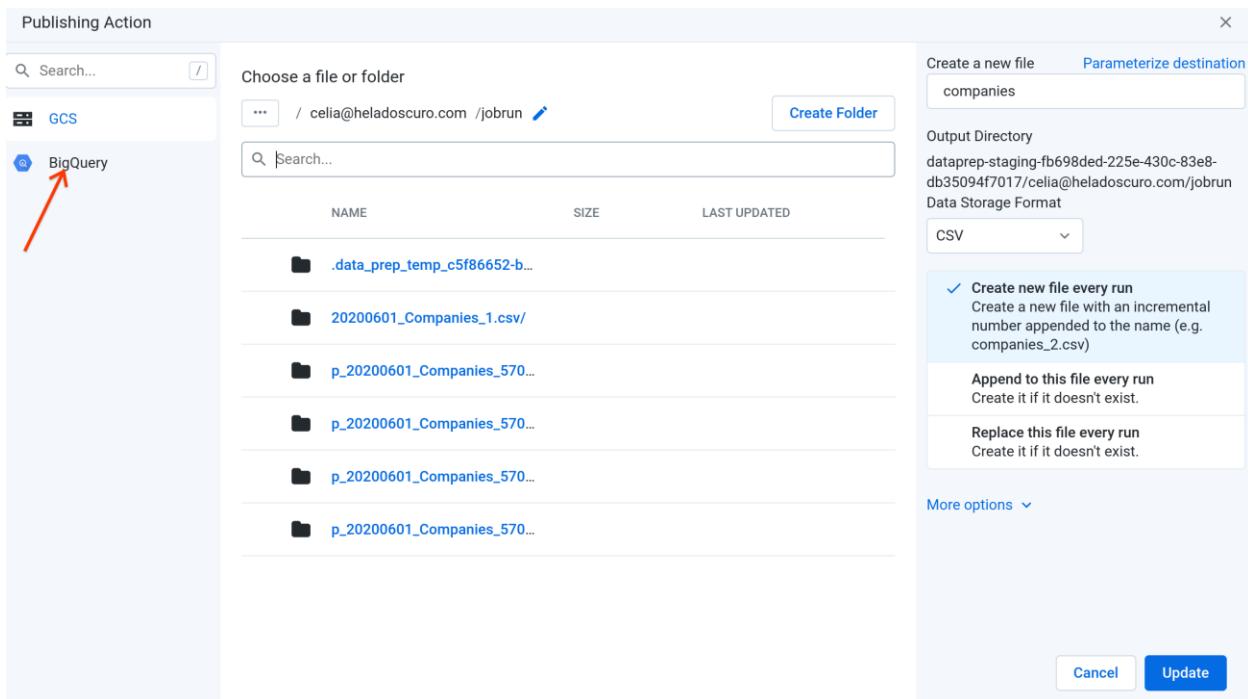
Cancel Run Job



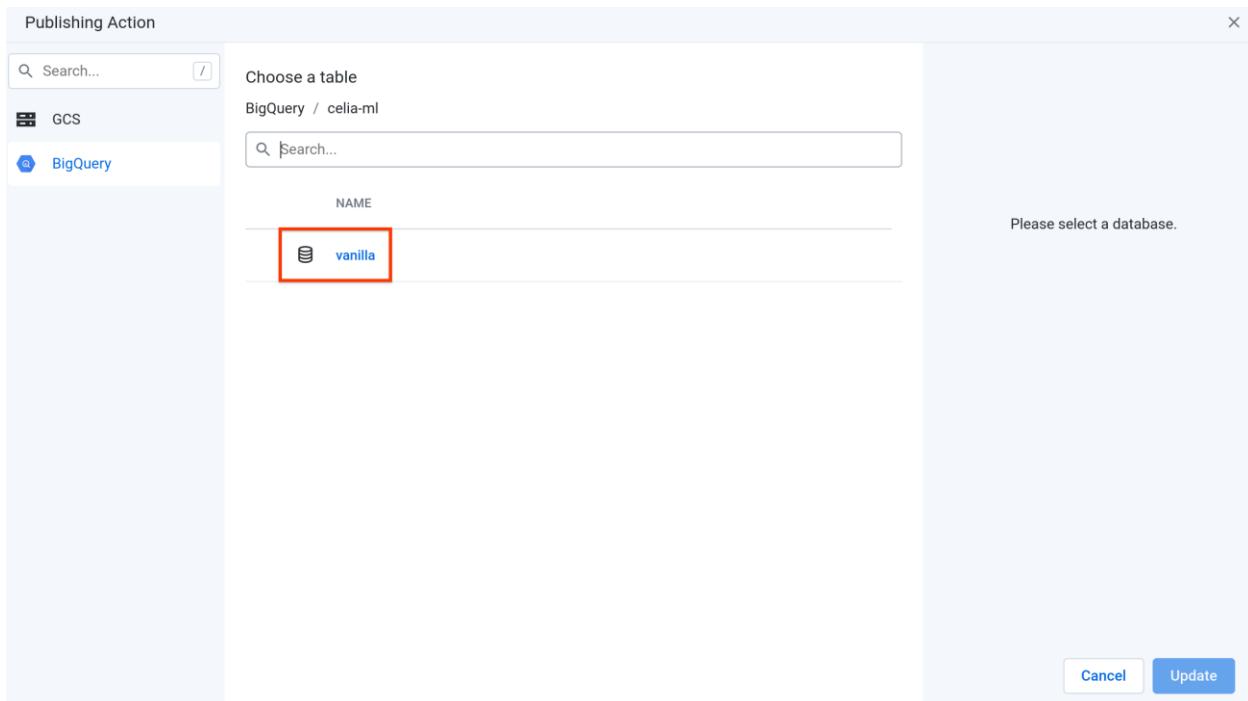
Inconsistent Company Names Demo

CELIA MURIEL

We select BigQuery



We click on our BigQuery dataset.



Inconsistent Company Names Demo

CELIA MURIEL

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](#)

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

CELIA MURIEL

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

We rename the CSV column names.

The screenshot shows the Google BigQuery interface with a table named 'companies'. The 'Company Name' column has a context menu open, with the 'Rename' option highlighted and circled in red. Other options in the menu include 'Change type', 'Move', 'Hide', 'Format', 'Calculate', 'Create column from examples', 'Group by', 'Pivot', 'Restructure', 'Filter rows', 'Replace', 'Standardize...', 'Extract', 'Split column', 'Column Details', and 'Show related Steps in Recipe'. The table has 8 columns, 49,386 rows, and 1 data type. The top navigation bar shows 'VANILLA FLOW > companies > Full Data' and includes 'Run Job' and other icons.



Inconsistent Company Names Demo

CELIA MURIEL

VANILLA FLOW > companies > Full Data

Preview

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	-
-	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.	-	TDD1284178Y8	-
-	MX	-	Tai Durango Cuatro, S.A.P.I. de C.V	-	TDC128417R14	-
-	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.	-	21-762787-0016	-
-	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

Show only affected Columns

Rename columns

Option: Manual rename required

Columns (1) Add

Company Name company_name

Add Cancel

VANILLA FLOW > companies > Full Data

New Step Recipe

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'

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		
-	-	-	-	Bucuresti Sector 2	28335	
-	-	-	-	Orkoien	31160	
-	-	-	-	Ciudad De Mexico	11550	
-	-	-	-	Ciudad De Mexico	11550	
-	-	-	-	Makati City	1200	
-	-	-	-	Zamudio	48170	
-	-	-	-	Zamudio	48170	
-	-	-	-	Montevideo	11200	
-	-	-	-	Madrid	28002	
-	-	-	-	Woking	GU21-6LQ	
-	-	-	-	Zamudio	48170	

8 Columns 49,386 Rows 1 Data Type

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



Inconsistent Company Names Demo

CELIA MURIEL

The screenshot shows a Dataflow job named 'companies' in the 'VANILLA FLOW' category. The job consists of 27 steps. The first few steps involve removing symbols from 'Town' and standardizing it. Subsequent steps rename columns and rows to follow a specific schema: 'company_name', 'country', 'id1', 'id2', 'id3', and 'id4'. The final step renames the ZIP code column to 'zipcode'.

Step	Action
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 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



Inconsistent Company Names Demo

CELIA MURIEL

Publishing Action

Choose a file or folder

GCS / celia@heladoscuro.com /jobrun

BigQuery

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

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

Publishing Action

Choose a table

BigQuery / celia-ml

BigQuery

vanilla

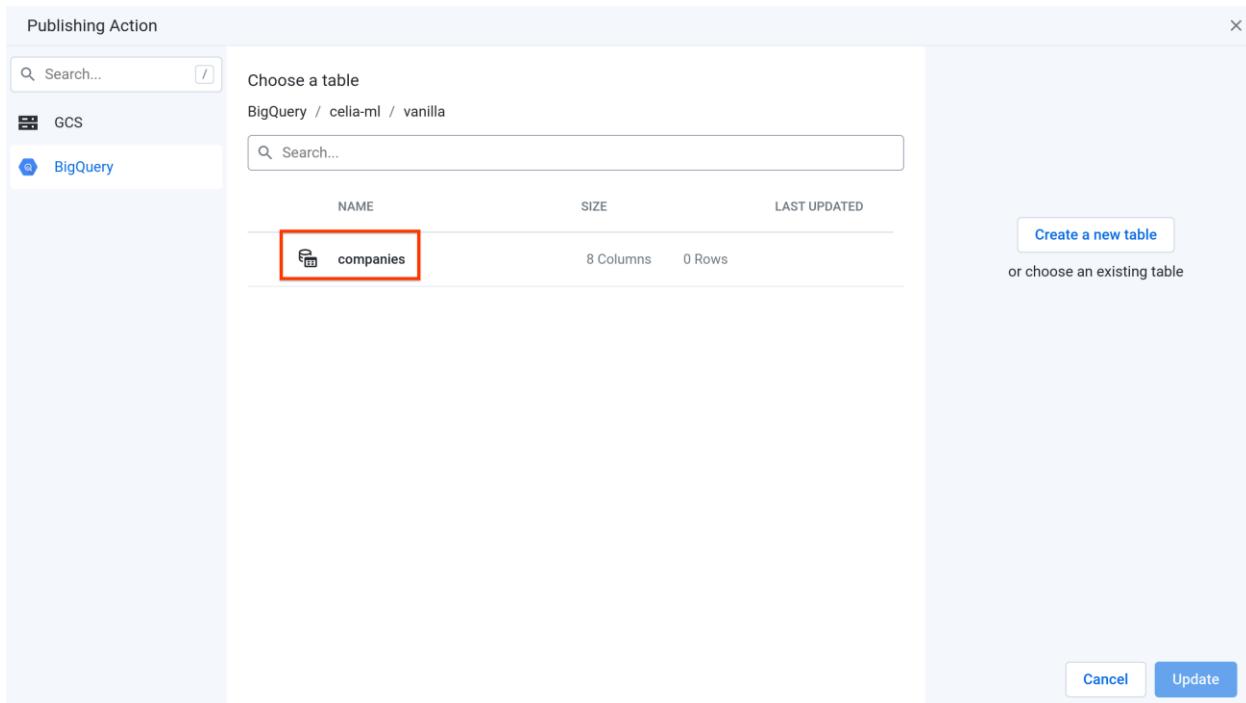
Please select a database.

Cancel Update



Inconsistent Company Names Demo

CELIA MURIEL



Publishing Action

Choose a table

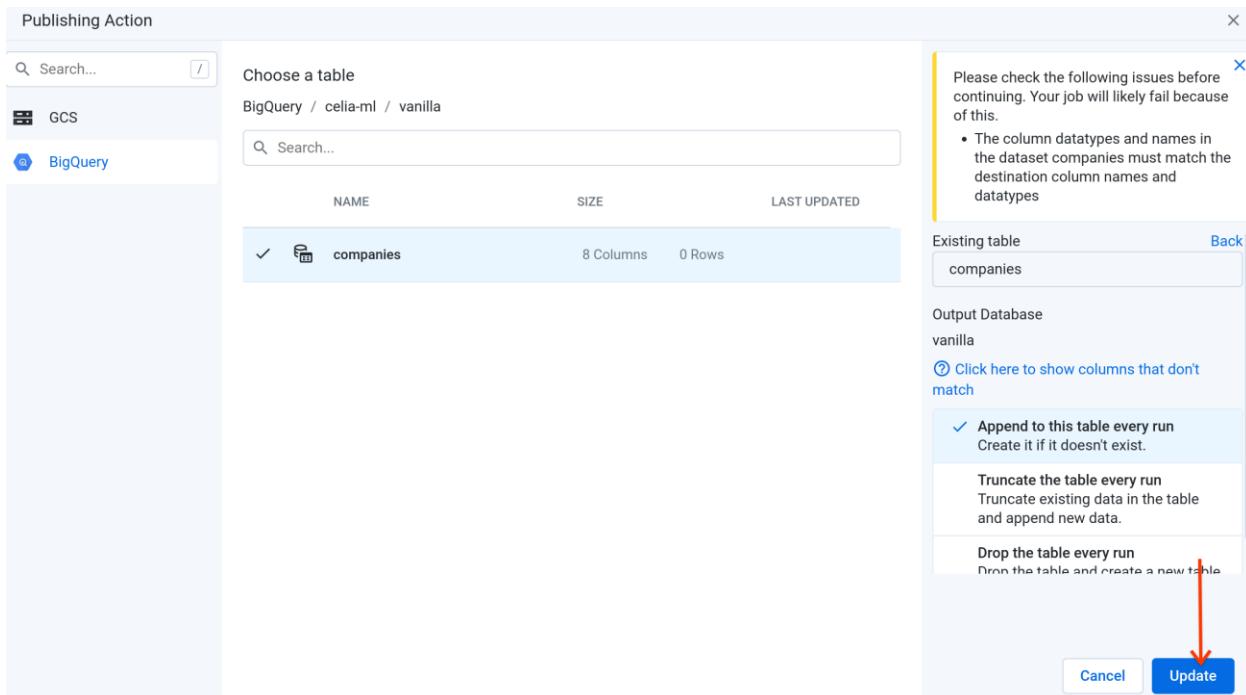
BigQuery / celia-ml / vanilla

NAME	SIZE	LAST UPDATED
 companies	8 Columns	0 Rows

Create a new table
or choose an existing table

Cancel Update

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



Publishing Action

Choose a table

BigQuery / celia-ml / vanilla

NAME	SIZE	LAST UPDATED
 companies	8 Columns	0 Rows

Please check the following issues before continuing. Your job will likely fail because of this.

- The column datatypes and names in the dataset companies must match the destination column names and datatypes

Existing table Back

Output Database

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

Click on “Run Job”.



Inconsistent Company Names Demo

CELIA MURIEL

Run Job on Dataflow

Options

Profile results
When enabled, this will generate a profile of your results

Publishing Actions

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

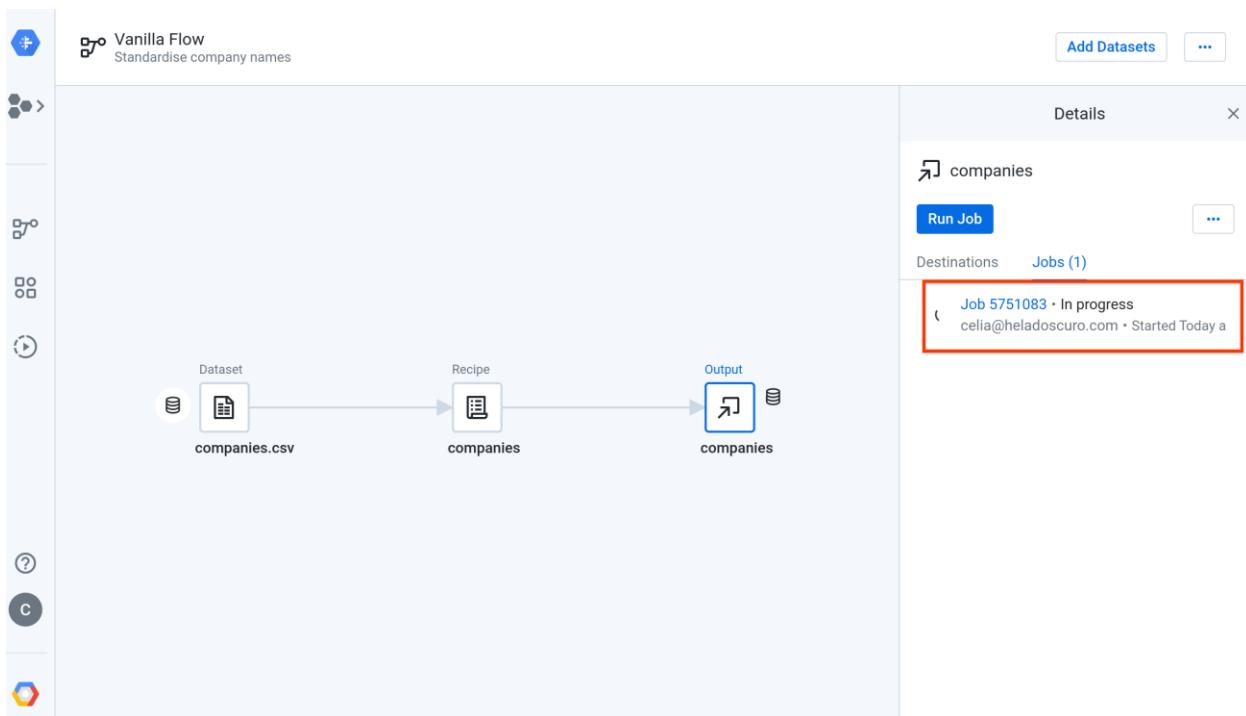
Machine Type: n1-standard-1

Advanced Settings

Add Publishing Action

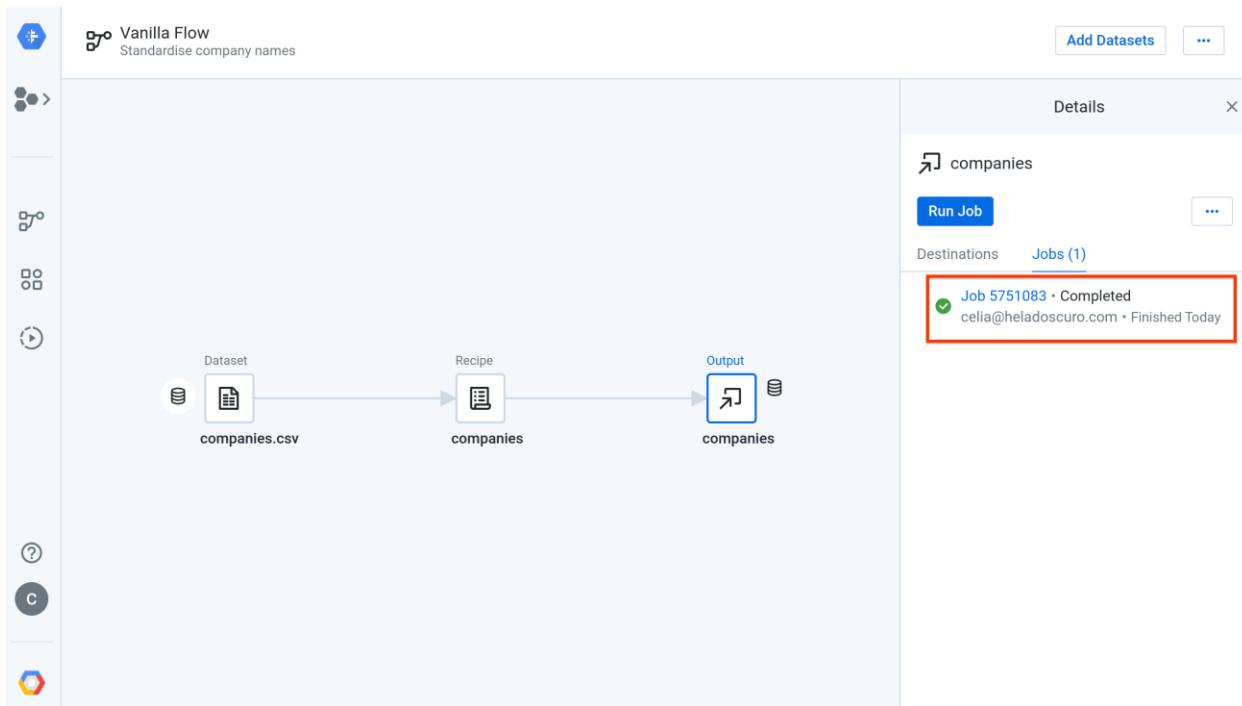
Cancel Run Job

Cloud Dataprep 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

CELIA MURIEL



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

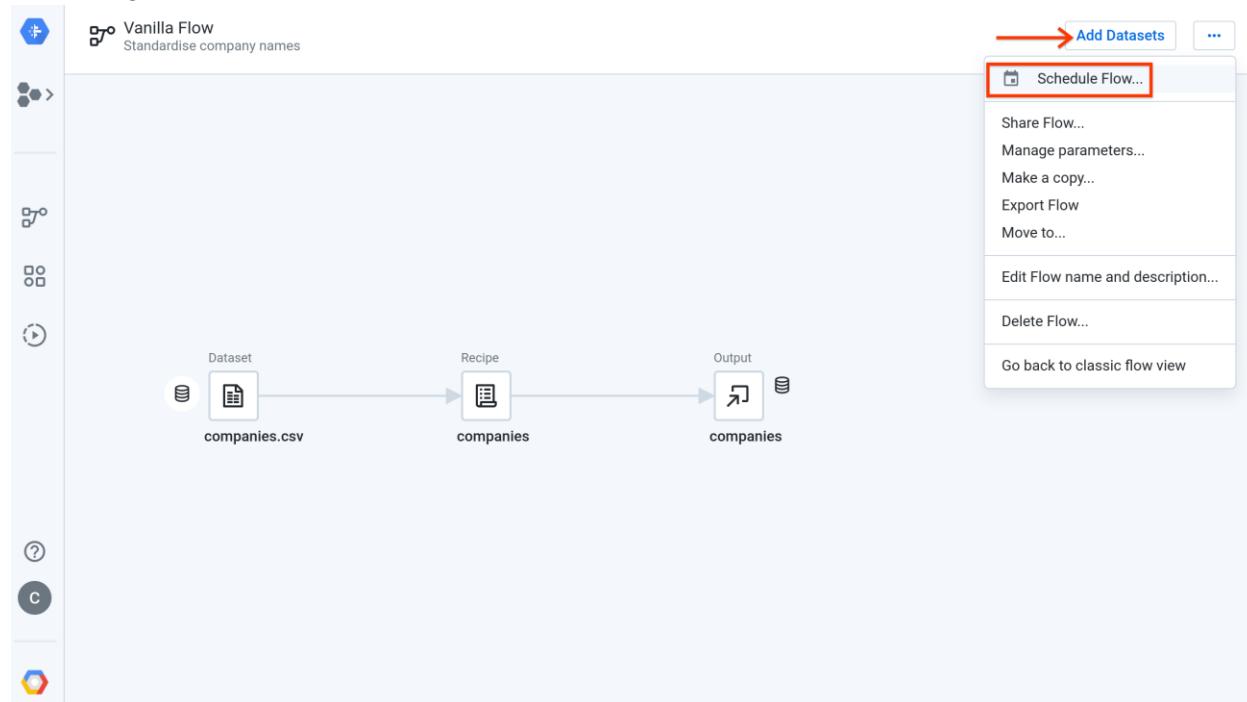
The screenshot shows the Google Cloud Platform BigQuery interface. The top navigation bar includes 'Google Cloud Platform', 'celia-ml', a search bar, and other navigation links. The main area has a sidebar with 'Query history', 'Saved queries', 'Job history', 'Transfers', 'Scheduled queries', 'Reservations', 'BI Engine', and 'Resources'. The 'Resources' section shows a dataset named 'celia-ml' containing a table named 'companies'. The 'Query editor' tab is active, displaying a simple SELECT query: 'select * from `vanilla.companies`;'. Below the editor, the 'Query results' section shows the output of the query. The results table has columns: Row, company_name, company_id1, company_id2, company_id3, company_id4, and country. The data rows are:

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, there are pagination controls: 'Rows per page: 100', '1 - 100 of 49386', 'First page', 'Last page', and arrows for navigating through the results.



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

CELIA MURIEL

VANILLA FLOW > companies > Full Data

New Step

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	Company Name	RBC	Company ID 1	RBC
		133 Categories	4,508 Categories	INDUSTRIAL GROUP LTD OF ASIA LTD	471 Categories
BR		Faculdade Leão Sampaio Fernando			No
BR		CPFL Renováveis			
PL		SIEMENS-GAMESA RENEWABLE ENERGY POLAND SP. ZOO			
PE		RED ELÉCTRICA ANDINA S.A.C. REA		28511261571	
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			
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			
LV		SIA MARINE SYSTEMS			
CA		Canadian Solar Solutions Inc.			
GB		British Solar Renewables Ltd.		159976146	
NL		NV Texels Eigen Stoomboot Ondernemi			
SV		Arturo Enrique Solano Urrutia-TECNO			

8 Columns 4,633 Rows 2 Data Types

VANILLA FLOW > companies > Full Data

Search Transformations

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	Company Name	RBC	Company ID 1	RBC
		133 Categories	4,508 Categories	INDUSTRIAL GROUP LTD OF ASIA LTD	471 Categories
BR		Faculdade Leão Sampaio Fernando			No
BR		CPFL Renováveis			
PL		SIEMENS-GAMESA RENEWABLE ENERGY POLAND SP. ZOO			
PE		RED ELÉCTRICA ANDINA S.A.C. REA		28511261571	
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			
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			
LV		SIA MARINE SYSTEMS			
CA		Canadian Solar Solutions Inc.			
GB		British Solar Renewables Ltd.		159976146	
NL		NV Texels Eigen Stoomboot Ondernemi			
SV		Arturo Enrique Solano Urrutia-TECNO			

8 Columns 4,633 Rows 2 Data Types



Inconsistent Company Names Demo

CELIA MURIEL

VANILLA FLOW > companies > Full Data

Search Transformations

Q do

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

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

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 conform to a regular expression.

8 Columns 4,633 Rows 2 Data Types

VANILLA FLOW > companies > Full Data

New formula

Formula type required

Single row formula

Create a new column from a single row formula

Formula required

DOUBLEMETAPHONEEQUALS()

DOUBLEMETAPHONEEQUALS(string1, string2, match_threshold)
Checks if two strings match phonetically using the Double Metaphone phonetic encoding algorithm. [Learn more](#)

Example
`DOUBLEMETAPHONEEQUALS(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](#)

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 Trifacta to use them.



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.

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

[Wrangle Language.](#) [Documentatio 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.

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



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

