

Quickstart

This page shows you how to get started using the Dataprep Web application.

Before you begin

Set up a project

1. [Sign in](https://accounts.google.com/Login) (https://accounts.google.com/Login) to your Google Account.

If you don't already have one, [sign up for a new account](https://accounts.google.com/SignUp) (https://accounts.google.com/SignUp).

2. In the Cloud Console, on the project selector page, select or create a Cloud project.

★ **Note:** If you don't plan to keep the resources that you create in this procedure, create a project instead of selecting an existing project. After you finish these steps, you can delete the project, removing all resources associated with the project.

[Go to the project selector page](https://console.cloud.google.com/projectselector2/home/dashboard) (https://console.cloud.google.com/projectselector2/home/dashboard)

3. Make sure that billing is enabled for your Google Cloud project. [Learn how to confirm billing is enabled for your project](/billing/docs/how-to/modify-project) (/billing/docs/how-to/modify-project).
4. Enable the Cloud Dataflow, BigQuery, and Cloud Storage APIs.

[Enable the APIs](https://console.cloud.google.com/flows/enableapi?apiid=dataflow.googleapis.com) (https://console.cloud.google.com/flows/enableapi?apiid=dataflow.googleapis.com)

Create a Cloud Storage bucket in your project

1. In the Cloud Console, go to the **Cloud Storage Browser** page.

[Go to the Cloud Storage Browser page](https://console.cloud.google.com/storage/browser) (https://console.cloud.google.com/storage/browser)

2. Click **Create bucket**.
3. In the **Create bucket** dialog, specify the following attributes:

- A unique bucket name, subject to the [bucket name requirements](/storage/docs/bucket-naming#requirements) (/storage/docs/bucket-naming#requirements).
- A [storage class](/storage/docs/storage-classes) (/storage/docs/storage-classes).
- A location where bucket data will be stored.

4. Click **Create**.

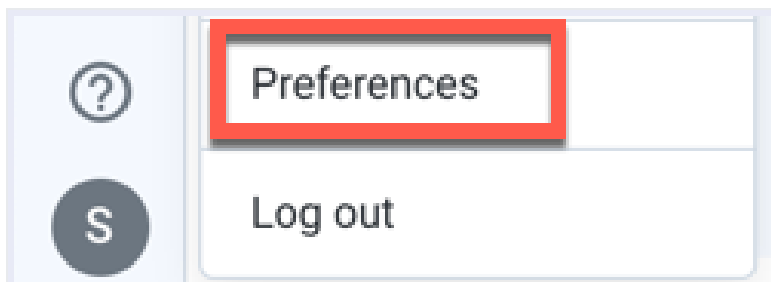
Set up Dataprep

Open [Dataprep](https://console.cloud.google.com/dataprep) (https://console.cloud.google.com/dataprep) on Google Cloud Console. When you first open Dataprep, the project owner is asked to allow data access by Google and Trifacta. The user must accept the terms of service, sign in to their Google account, and choose a Cloud Storage bucket to use with Dataprep (see [Enabling Dataprep](/dataprep/docs/resources/enable-disable#enabling) (/dataprep/docs/resources/enable-disable#enabling)).

After completing these steps, the Dataprep home page appears. You can choose to Show the tour and run it, which walks you through steps that parallel the steps in this quickstart.

The screenshot shows the Google Cloud Dataprep interface. At the top, it says "Cloud Dataprep by TRIFACTA". Below that, a welcome message "Welcome back, S!" is displayed. The main area features an illustration of three people (two men and one woman) interacting with data flows and documents. A "Start tour" button is prominently displayed in the center. To the right, there are buttons for "Import Data" and "Create Flow". A sidebar on the right lists various resources like "What's New", "Release Notes", "Newsletter", "Tutorial Video", "Articles", "Community", and "Wrangle Exchange". At the bottom, there's a "Recently Updated" section showing a dataset "[2f7ca1a0] Dataset with Parameters Flow" updated on "01/30/2020".

How to change Dataprep bucket directories. You can change the Cloud Storage bucket directory (folders) used by Dataprep for uploads, job runs, and temp storage by selecting the User icon in the left pane, which displays the first initial of the account owner's username. Then select **Preferences**.



The User Account settings page opens, where you can change each of the bucket directory settings.

The 'User Account' settings page is displayed. At the top, there is a circular profile icon with the letter 'U' and the text 'User Account'. Below this, there are four sections for editing user information and directory settings. Each section has a text input field and a 'Change' button. The 'Name' field contains 'User Name'. The 'Email' field contains 'username@domain.com'. The 'Upload directory', 'Job Run directory', and 'Temp directory' fields all contain the same Google Cloud Storage path: 'gs://dataprep-staging-783b8924-ca50-4da4-b354-2091523b...'. A blue 'Done' button is located at the bottom right of the settings area.

Create a flow

Dataprep uses a container object called a `flow` to access and manipulate datasets. From the Dataprep home page, click the Flows icon



in the left nav bar. Then, click **Create**. Select **Create Flow**. Fill in a flow name and description, then click **Create**. Since this quickstart uses 2016 data from the United States Federal Elections Commission (<https://www.fec.gov/>), you may wish to name it, "FEC-2016", with a description that refers to this data.

Create Flow ✕

Flow Name

Flow Description

Cancel Create

The Flow View page opens.

FEC-2016 quickstart ⋮

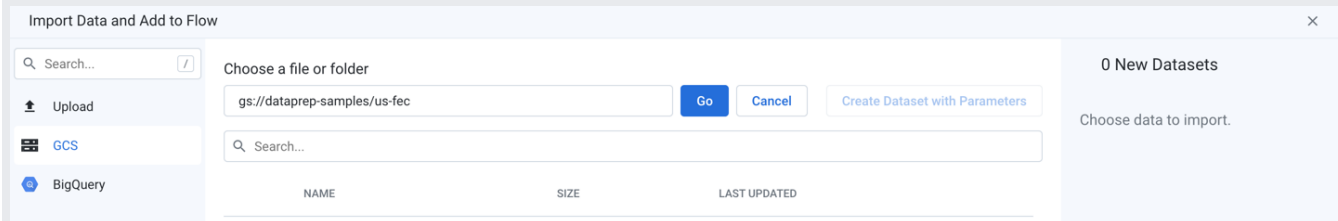
A square icon with a document symbol and a grid pattern, representing a dataset.

Add Datasets into this Flow to start wrangling.

[Add Datasets](#)


Import datasets

From Flow View, click **Add Datasets** to open the Add Datasets to Flow page. Click **Import Datasets**. Select **GCS** in the left panel. Under **Choose a file or folder**, click the Pencil icon, then insert `gs://dataprep-samples/us-fec` in the GCS text box. Click **Go**.



Select the `cn-2016.txt` dataset. Name it "Candidate Master 2016" in the right panel. Then select the `itcont-2016.txt` dataset, naming it "Campaign Contributions 2016". After both datasets are listed and renamed in the right panel, click **Import & Add to Flow** to add the datasets.


2 New Datasets Clear All

 Campaign Contributions 2016 ×

Add a Description

ABC column2	ABC column3	ABC column4
C00000935	A	M10
C00000935	A	M4
C00000935	A	M6
C00000935	A	M7
C00000935	A	M8

[Edit settings](#)

 Candidate Master 2016 ×

Add a Description

ABC column2	ABC	column3
H0AK00097	COX, JOHN R.	
H0AL02087	ROBY, MARTHA	
H0AL02095	JOHN, ROBERT E JR	
H0AL05049	CRAMER, ROBERT E	

[Import & Add to Flow](#) [Cancel](#)

Wrangle the Candidate file

On the FEC 2016 Flow page, select the Candidate Master 2016 dataset, then click **Add New Recipe**.

The screenshot shows the Google Cloud Dataprep interface for a project named "FEC-2016" (United States Federal Elections Commission 2016). In the main workspace, a dataset "Candidate Master 2016" is visible. A blue "Add New Recipe" button is positioned above it. Below the dataset, there is a "Data Preview" section showing a table of data. To the right, a "Details" panel is open, displaying information about the dataset.

ABC column2	ABC	column3	ABC col
H0AK00097	COX, JOHN R.		REP
H0AL02087	ROBY, MARTHA		REP
H0AL02095	JOHN, ROBERT E JR		IND
H0AL05049	CRAMER, ROBERT E "BUD" JR		DEM
H0AL05163	BROOKS, MO		REP
H0AL06088	COOKE, STANLEY KYLE		REP
H0AL07086	SEWELL, TERRYCINA ANDREA		DEM

Details panel information:

- Type: GCS
- Location: gs://dataprep-samples/us-fec/cn-2016.txt
- File Size: 717.52kB
- Size: 15 columns · 4 types
- Updated: Today at 4:32 PM
- Created: Today at 4:32 PM

A new recipe icon appears. Click **Edit Recipe**.

The screenshot shows the Google Cloud Dataprep interface after a new recipe has been created. The main workspace now displays two recipe icons: "Candidate Master 2016" and "Candidate Master 2016 - 2". The "Candidate Master 2016 - 2" icon is highlighted with a blue circle and a plus sign, indicating it is the active recipe. The "Details" panel on the right is updated to show information for "Candidate Master 2016 - 2".

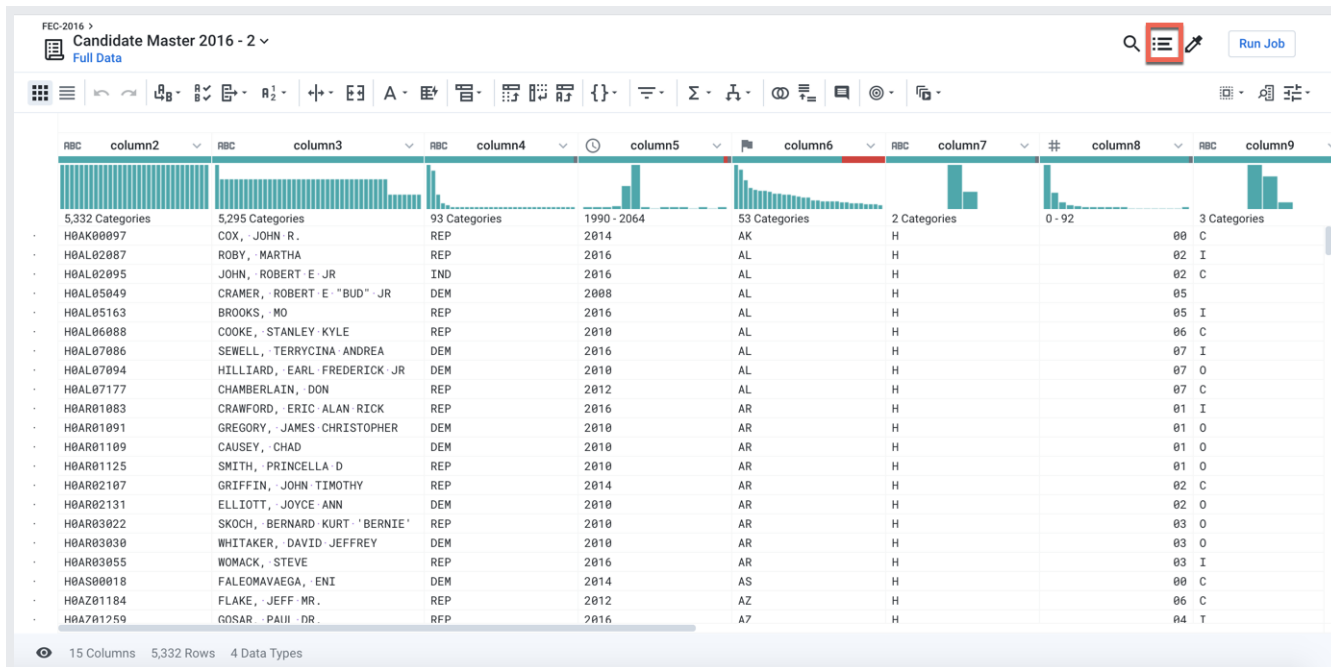
Details panel information:

- Recipe: Data
- Steps Preview: (empty)

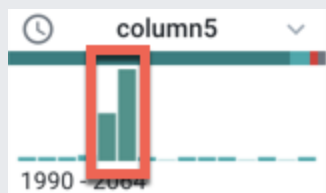
The Transformer page opens, where you can explore a sample of your data and build your recipe by applying transformation steps to it.

To Display the Recipe pane: A recipe is created in the Recipe pane. If the Recipe pane is not displayed on the right side of the page, click the Recipe icon at the top-right of the Grid view page.

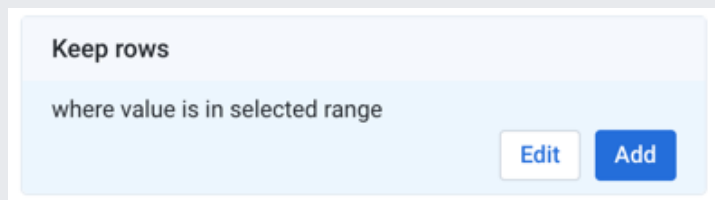




column5 (Date/Time) contains a year value. Select the years 2016 and 2017 in the histogram by dragging across them.



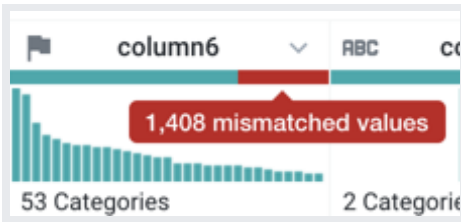
In the right panel, you should see a suggestion card titled "Keep rows where value is in selected range". Click **Add**.



The following recipe step is added to the recipe:

```
Keep rows where(date(2016, 1, 1) <= column5) && (column5 < date(2018, 1, 1))
```

In the column6 (State) header, hover over and click the mismatched (red) bar to select the mismatched rows.



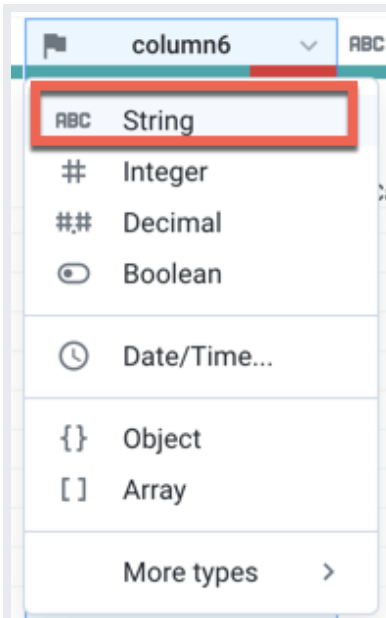
In the Status bar at the bottom of the page, select the Show only affected checkbox. Notice that some the red highlighted (mismatched) items have the value "US" in column6 and "P" in column7. These are presidential candidates. The mismatch occurs because column6 is marked as a "State" column (indicated by the flag icon), but it also includes non-state (such as "US") values.

column6	column7
AS	H
GU	H
MP	H
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P
US	P

Show only affected Rows

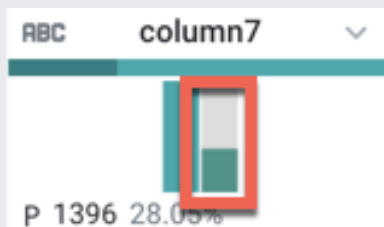
d

To correct the mismatch, click the X in the right panel to cancel the transformation. The column must be re-typed as a column of String data type. Click the flag icon above column6 and select "String".

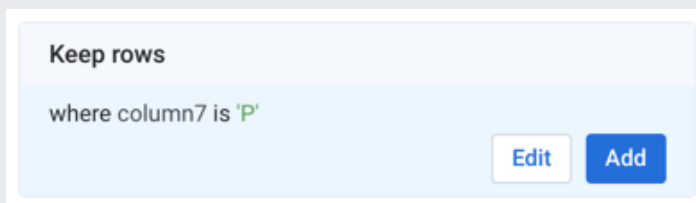


The column's data type is changed to String. String data type matches with any non-empty value in a column, which means that the mismatch is removed. The data quality bar is now completely green.

Now, let's filter on just the presidential candidates. In the histogram for column7, click the "P" bin.

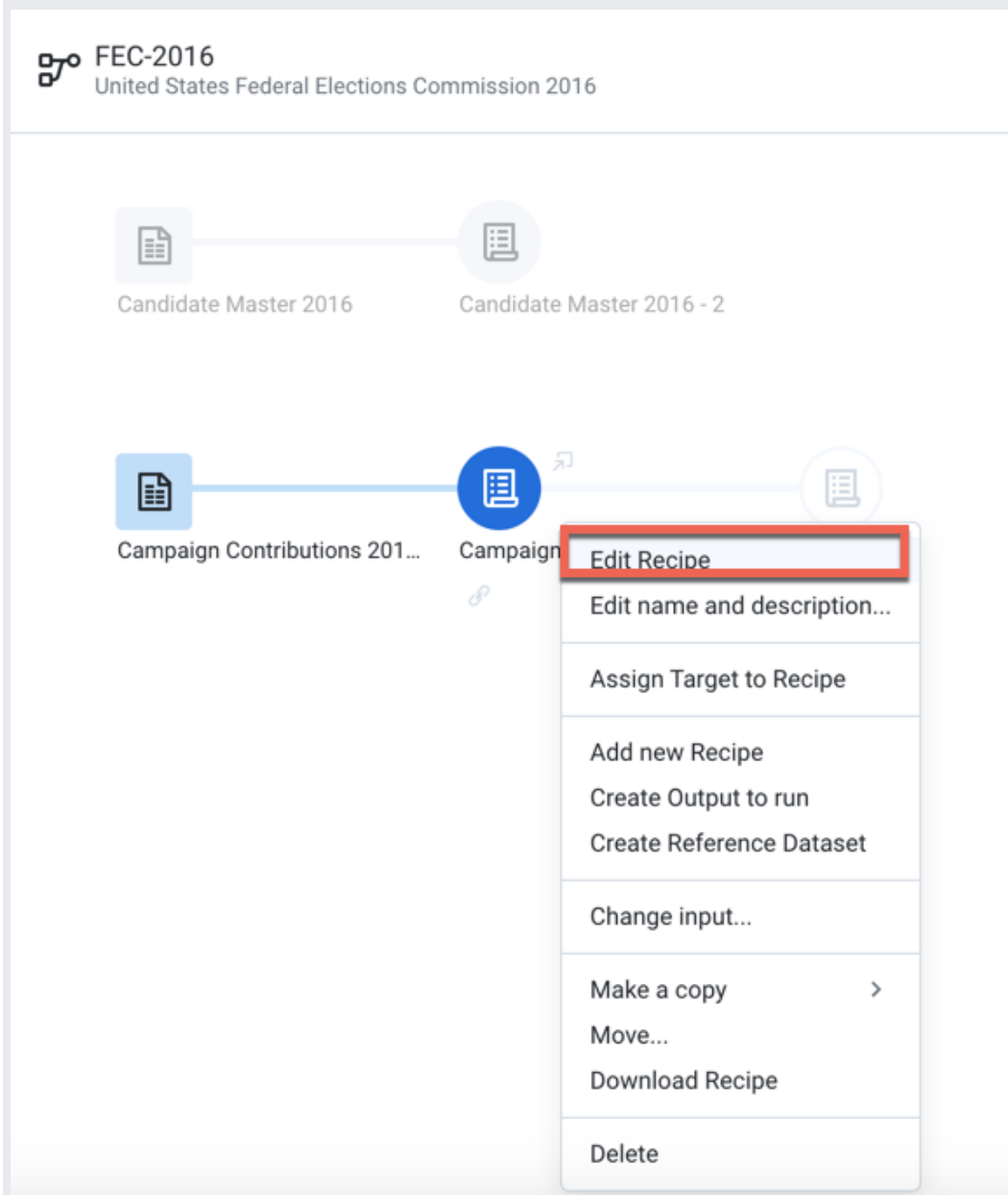


In the right panel, you should see a suggestion card titled "Keep rows where column7 is 'P':. Click **Add**.



Wrangle the Contributions file and join it in

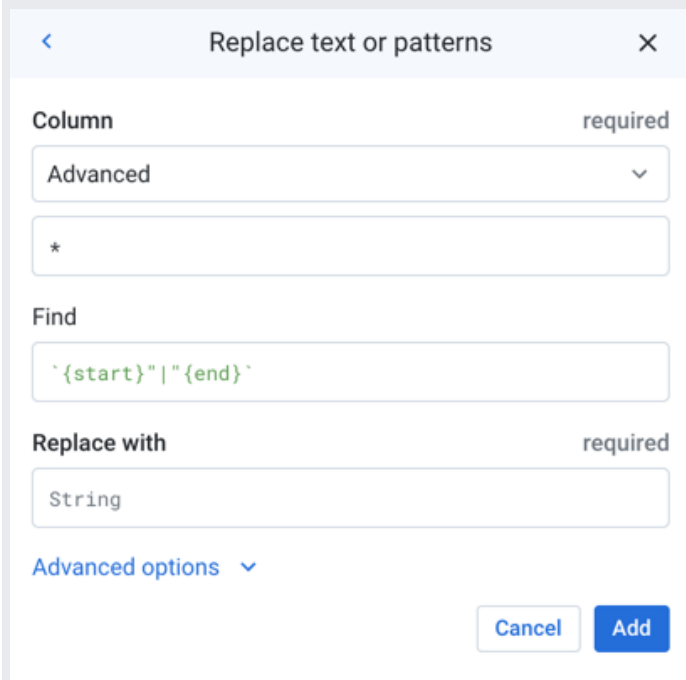
In the Transformer page, click **FEC 2016** in the upper left corner of the Flow View page. Select **Campaign Contributions 2016**, then select **Add new Recipe**, then click **Edit Recipe** to open a sample of the contributions dataset back in the Transformer page.



In the Transformer page, open the Recipe panel. You can add a new step to the recipe that removes extra delimiters from the contributions dataset. Open the Recipe panel. Copy and paste the following Wrangle language command in the Search box.

```
replacepatterns col: * with: ' ' on: `{start}"|"{end}` global: true
```

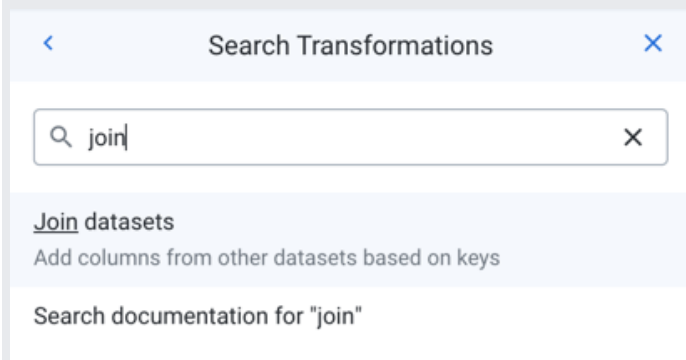
The Transformation Builder parses the Wrangle command and populates the fields for the Replace text or patterns transformation. Click **Add** to add the transformation to the recipe.



The screenshot shows a dialog box titled "Replace text or patterns" with a close button (X) in the top right corner. The dialog contains the following fields and controls:

- Column:** A dropdown menu with "Advanced" selected and a "required" label to its right.
- Find:** A text input field containing the regular expression ``{start}"|"{end}``.
- Replace with:** A text input field containing the word "String" and a "required" label to its right.
- Advanced options:** A blue link with a downward arrow.
- Buttons:** "Cancel" and "Add" buttons at the bottom right.

Joins: Let's join this dataset to the first one. Add another new step to the recipe, then type "Join" in the Search box.



The screenshot shows a dialog box titled "Search Transformations" with a close button (X) in the top right corner. The dialog contains the following elements:

- Search Input:** A search box with a magnifying glass icon, containing the text "join" and a close button (X).
- Search Results:**
 - Join datasets:** A link with a blue underline, followed by the text "Add columns from other datasets based on keys".
 - Search documentation for "join":** A text link.

Click the "Join datasets" link to open the Joins page.

In the Recipes in the current flow tab, select **Candidate Master 2016-2** (the Candidate Master dataset with the Recipe steps added above).

Choose dataset or recipe to join with Campaign Contributions 2016
- 4

Search...

Recipes in current flow Datasets in current flow All datasets

Name	Last Updated	Source	Data	Recipe
Campaign Contributions 20...	Today at 5:00 PM	(this flow)		
<input checked="" type="checkbox"/> Candidate Master 2016 - 2	Today at 4:54 PM	(this flow)		

No records to display. Edit your Recipe to view a larger sample in Transformer.

Browse current flow

Cancel Accept

Click **Accept**.

In the Join window, you specify the keys and conditions of the join transformation. Cloud Dataprep infers some of these details for you. In this case, the join type (inner) is correct, but the join keys are not. Hover over them and click the Pencil icon. Select **Current = column 2** and **Joined-in = column 11** as the join keys.

Current [?] required

column2 X v

Joined-in [?] required

column11 X v

Fuzzy match

Click **Save and Continue**. Click **Next**. In the Join - Output Columns window, select the checkbox immediately under the "All (36)" label, which adds all columns of both datasets to the joined dataset:

All (36) Current (21) Joined-In (15)

<input checked="" type="checkbox"/>	Column	Source
<input checked="" type="checkbox"/>	column2	Current
<input checked="" type="checkbox"/>	column11	Joined-In

Then click **Review**. If all looks good, click **Add to Recipe**. In the Transformer page, the join transformation has been applied.

Create a summary: Add the following steps to the recipe to generate a useful summary by aggregating, averaging and counting the contributions in column 16 and grouping on the candidates by IDs, names, and party affiliation in columns 2, 9, 8 respectively. Click **New Step** in the Recipe panel. Then, copy the following step and paste it into the Search box:

```
pivot value:sum(column16),average(column16),countif(column16 > 0) group: column2,col
```

The screenshot shows a 'Pivot columns' dialog box with the following configuration:

- Column labels:** Select column(s)
- Row labels:** RBC column2, RBC column9, RBC column8
- Values:** SUM(column16), AVERAGE(column16), COUNTIF(column16 > 0)

Buttons: Cancel, Add

A sample of the joined and aggregated data is displayed, representing a summary table of US presidential candidates and their 2016 campaign contribution metrics.

FEC-2016 > Campaign Contributions 2016 - 4 v
Initial Sample

6 Columns 52,533 Rows 3 Data Types

RBC	column2	RBC	column9	RBC	column8	#	sum_column16	##	average_column16	#	countif
	2,685 Categories		50,762 Categories		6 Categories		-5.4k - 2.02M		-5,400 - 1,011,250		0 - 21
-	C00422410		PIRILLO, CAROLYN		IND			25		25	1
-	C00425470		SMITH, ALICIA W.		IND			250		250	1
-	C00431056		CARDINALE, GERALD P		IND			1000		1000	1
-	C00435321		MADRID, SYLVIA		IND			28		28	1
-	C00436550		SPENCER, BILL DR.		IND			250		250	1
-	C00438655		NYQUIST, LAURA K		IND			62		62	1
-	C00442905		HEIMBUCK, MICHAEL JOEL		IND			2500		2500	1
-	C00448696		JACK, CAROL C.		IND			50		50	1
-	C00456335		WELD, SUSAN ROOSEVELT		IND			3		3	1
-	C00458158		FOWLER, JOHN ELLIOTT MR.		IND			75		75	1
-	C00458844		BATISTA, LUIS		IND			20		20	1
-	C00459933		JONES, EDDIE J		IND			10		10	1
-	C00469205		LEVINE, PETER		IND			100		100	1
-	C00473827		TRUMP, JULIUS MR		IND			1000		1000	1
-	C00475350		SPENCER, STEPHEN S		IND			47		47	1
-	C00487447		BREHM, TERRY LEE		IND			77		77	1
-	C00489419		GRECO, FRANK DR.		IND			85		85	1
-	C00493304		OCASIO, JOSE L. JR.		IND			10		10	1
-	C00495028		MCCART, ALLEN		IND			10		10	1
-	C00496307		BROWN, MATTHEW		IND			11		11	1
-	C00496307		CARTER, JONATHAN		IND			12		12	1

You can make the data easier to interpret by adding the following renaming and rounding steps to the recipe.

```
rename type: manual mapping: [column9, 'Candidate_Name'],
[column2, 'Candidate_ID'], [column8, 'Party_Affiliation'],
[sum_column16, 'Total_Contribution_Sum'],
[average_column16, 'Average_Contribution_Sum'],
[countif, 'Number_of_Contributions']
```

```
set col: Average_Contribution_Sum value: round(Average_Contribution_Sum)
```

FEC-2016 > Campaign Contributions 2016 - 4 v
Initial Sample

6 Columns 52,533 Rows 3 Data Types

RBC	Candidate_ID	RBC	Candidate_Name	RBC	Party_Affiliation	#	Total_Contribution_Sum	#	Average_Contribution_Sum	#	
	2,685 Categories		50,762 Categories		6 Categories		-5.4k - 2.02M		-5.4k - 1.01M		0 - 21
-	C00422410		PIRILLO, CAROLYN		IND			25		25	
-	C00425470		SMITH, ALICIA W.		IND			250		250	
-	C00431056		CARDINALE, GERALD P		IND			1000		1000	
-	C00435321		MADRID, SYLVIA		IND			28		28	
-	C00436550		SPENCER, BILL DR.		IND			250		250	
-	C00438655		NYQUIST, LAURA K		IND			62		62	

Gather new samples

As needed, you can generate a different kind of sample of your data, which helps to locate outliers and to verify that your transformation steps apply to all rows in the dataset. To view more data, click the "Initial Sample" link at the top-left of the page. In the Samples pane, select a random, quick sample, then click **Collect**.

The screenshot shows the 'Samples' pane with the following content:

- Current Sample**
 - Initial** 56,384 rows
 - Today at 4:55 PM
 - [See all collected samples...](#)
- Collect New Sample**
 - Recently collected
 - New Random sample** (Quick scan. Today at 4:58 PM) with a **Load sample** button.
 - Select a method to collect a new sample
 - Random**: Randomly select rows from the dataset.
 - First Rows**: Select the first rows from the dataset.

In the Samples pane, select a random sample. For the scan type, select Quick. You can name the sample if needed. Then, click **Collect**.

< Collect new sample X

Name

Scan required

Quick

Full

After the job completes, click **Load Sample** in the Samples panel to load the new sample into the Transformer page.

FEC-2016 >

Campaign Contributions 2016 - 4 Random Run Job

6 Columns 42,587 Rows 2 Data Types

RBC	Candidate_ID	RBC	Candidate_Name	RBC	Party_Affiliation	#	Total_Contribution_Sum	#	Average_Contribution_Sum	#
2,517 Categories		41,095 Categories		6 Categories			-5k - 1.5M		-5k - 1.5M	0 - 14
-	C00165282	-	OREAR, NORMAN	-	IND	-		-		9
-	C00127779	-	DURNING, DAVID M.	-	IND	-		-		96
-	C00116632	-	DEATON, CHRISTOPHER D.	-	IND	-		-		500
-	C00197228	-	DOLT, HOWARD	-	IND	-		-		16
-	C00163832	-	JACKSON, ARTHUR	-	IND	-		-		94
-	C00577130	-	FLEITAS, VICTOR	-	IND	-		-		33
-	C00590778	-	LOWRY, PAMELA L.	-	IND	-		-		500
-	C00482984	-	NIEDFELDT, LAUREEN R MS	-	IND	-		-		25
-	C00575795	-	STORNELLO, MICHAEL	-	IND	-		-		250
-	C00000935	-	SACKLER, JESSIE B.	-	IND	-		-		75
-	C00255752	-	JANOSY, NORAH R. M.D.	-	IND	-		-		50
-	C00575795	-	RYAN, JOANNA	-	IND	-		-		5
-	C00574624	-	BRANDT, DAVID	-	IND	-		-		35
-	C00000901	-	BANKS, ROSLYN	-	IND	-		-		14
-	C00035519	-	TITSWORTH, STACI LEIGH	-	IND	-		-		20
-	C00572966	-	DEMOCRATIC LEGISLATIVE CAMP	-	ORG	-		-		19626
-	C00401224	-	LAMB, BRENDON	-	IND	-		-		50
-	C00076810	-	KELLY, KEVIN S	-	IND	-		-		15
-	C00465492	-	OATIS, CAROL A	-	IND	-		-		25
-	C00575795	-	WEXLER, CYNTHIA	-	IND	-		-		375
-	C00586537	-	GANDY, ISAAC	-	IND	-		-		50

Run a job

You can now run a job to apply your changes across the entire joined dataset. In the Transformer page, click **Run Job**.

In the Run Job on Dataflow page:

- Select the Profile Results checkbox. When a profile is generated, you can review a statistical and visual summary of the results of your job, which is useful for evaluating the quality of your transformations.
- By default, a CSV file is generated with your job. Suppose you want to add a JSON output file, too. Click **Add Publishing Action**.
 1. Click **Create a new file**.
 2. Specify a new name for the file if desired.
 3. For the Data Storage Format, select **JSON** from the drop-down.
 4. You can explore the other options if you want. Click **Add**. The publishing action is added to your job specification.
- Click **Run Job**.
- The job is queued for execution in Cloud Dataflow.

Tracking progress: In Flow View, you can see the progress of the job in the right panel. To explore details, click the Job Id link.

Profile: When the job completes successfully, click the Profile tab to see the visual profile of your job results.

Export Results: Click the Output Destinations tab. The output files are listed. From a file's context menu, you can click **View on Google Cloud Storage**. You can download from there.

What's next

- [View the Dataprep video presentation](https://www.youtube.com/watch?v=Q5GuTlgmt98) (https://www.youtube.com/watch?v=Q5GuTlgmt98).

Except as otherwise noted, the content of this page is licensed under the [Creative Commons Attribution 4.0 License](https://creativecommons.org/licenses/by/4.0/) (https://creativecommons.org/licenses/by/4.0/), and code samples are licensed under the [Apache 2.0 License](https://www.apache.org/licenses/LICENSE-2.0) (https://www.apache.org/licenses/LICENSE-2.0). For details, see the [Google Developers Site Policies](https://developers.google.com/site-policies) (https://developers.google.com/site-policies). Java is a registered trademark of Oracle and/or its affiliates.

Last updated 2020-06-23 UTC.