Six Core Data Wrangling Activities

An introductory guide to data wrangling with Trifacta

Today’s Data Driven Culture

Are you inundated with data?

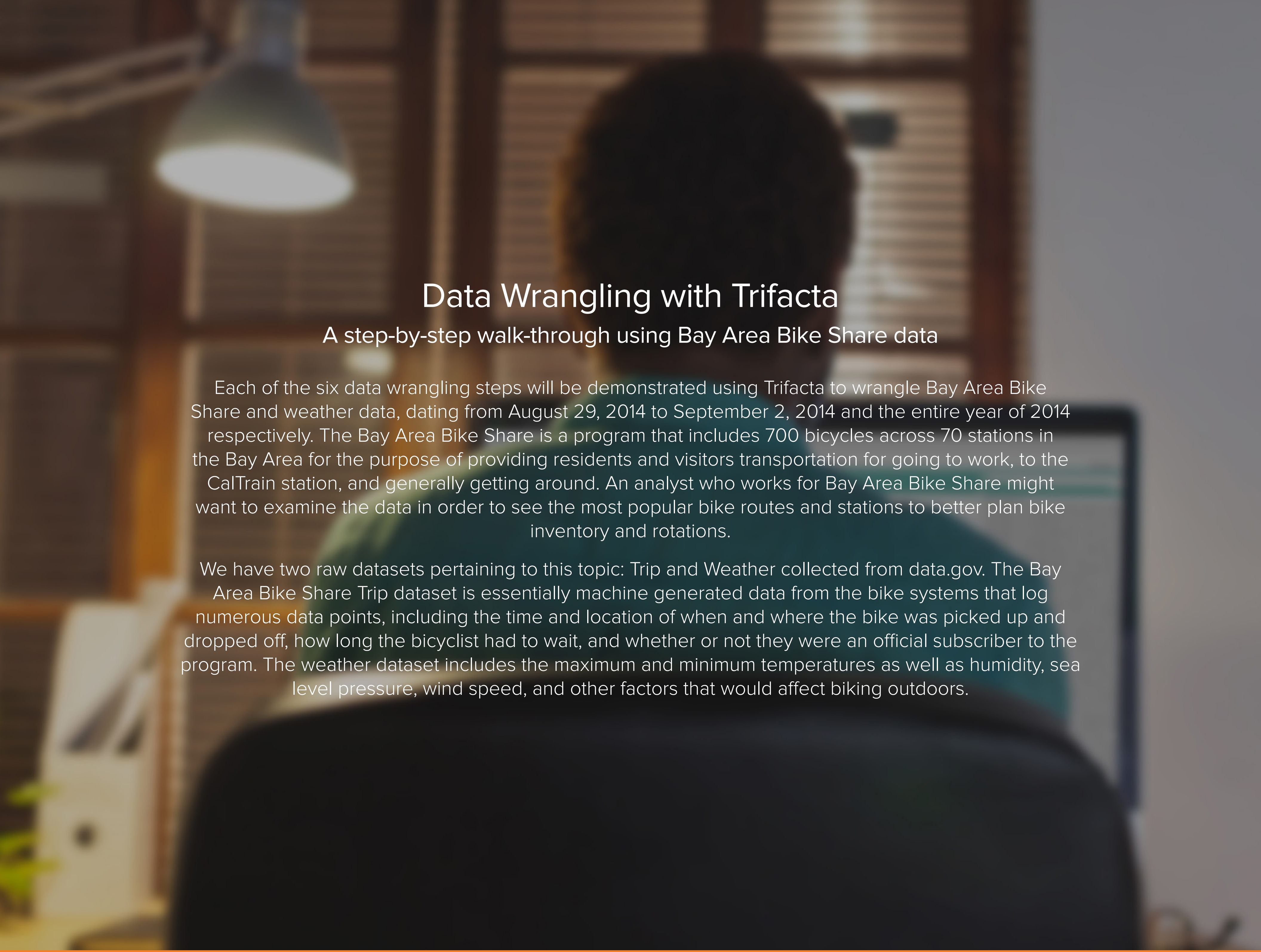
Today, most organizations are collecting as much data in one week as they used to collect in one year. Even though they have more data, they’re not using the bulk of it —most companies analyze less than **20%** of their data. Why? Preparing diverse datasets for analysis is time-consuming and messy, often left to data engineers or data scientists.

Introducing, Trifacta Wrangler!

With Trifacta Wrangler, analysts, researchers and data visualization specialists no longer need to be dependent on IT for data preparation. Trifacta empowers you to wrangle data yourself—allowing you to spend less time waiting and more time on your analysis.

Start wrangling your data by following the 6 steps of the process:

- Discovering
- Structuring
- Cleaning
- Enriching
- Validating
- Publishing

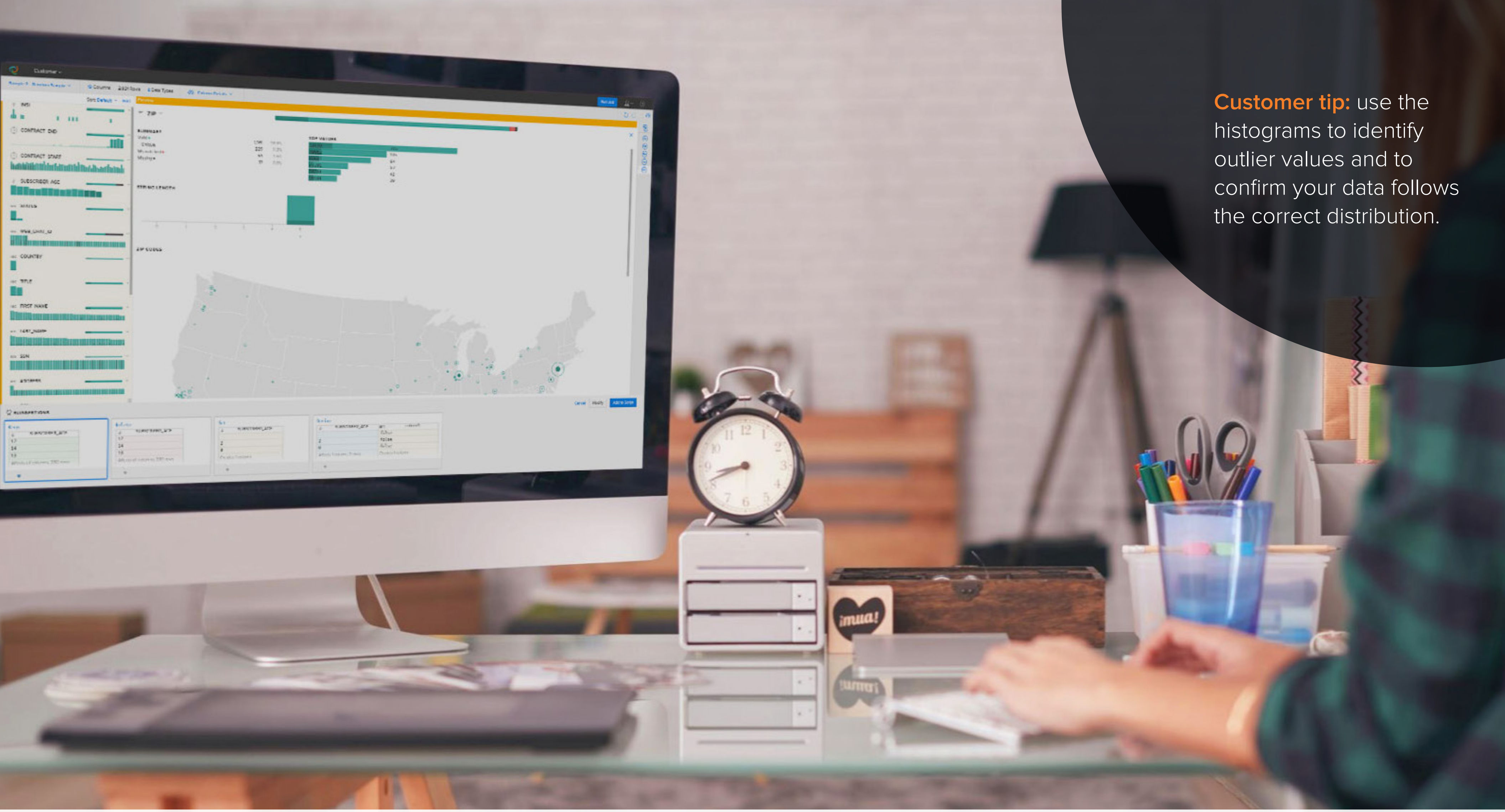


Data Wrangling with Trifacta

A step-by-step walk-through using Bay Area Bike Share data

Each of the six data wrangling steps will be demonstrated using Trifacta to wrangle Bay Area Bike Share and weather data, dating from August 29, 2014 to September 2, 2014 and the entire year of 2014 respectively. The Bay Area Bike Share is a program that includes 700 bicycles across 70 stations in the Bay Area for the purpose of providing residents and visitors transportation for going to work, to the CalTrain station, and generally getting around. An analyst who works for Bay Area Bike Share might want to examine the data in order to see the most popular bike routes and stations to better plan bike inventory and rotations.

We have two raw datasets pertaining to this topic: Trip and Weather collected from data.gov. The Bay Area Bike Share Trip dataset is essentially machine generated data from the bike systems that log numerous data points, including the time and location of when and where the bike was picked up and dropped off, how long the bicyclist had to wait, and whether or not they were an official subscriber to the program. The weather dataset includes the maximum and minimum temperatures as well as humidity, sea level pressure, wind speed, and other factors that would affect biking outdoors.



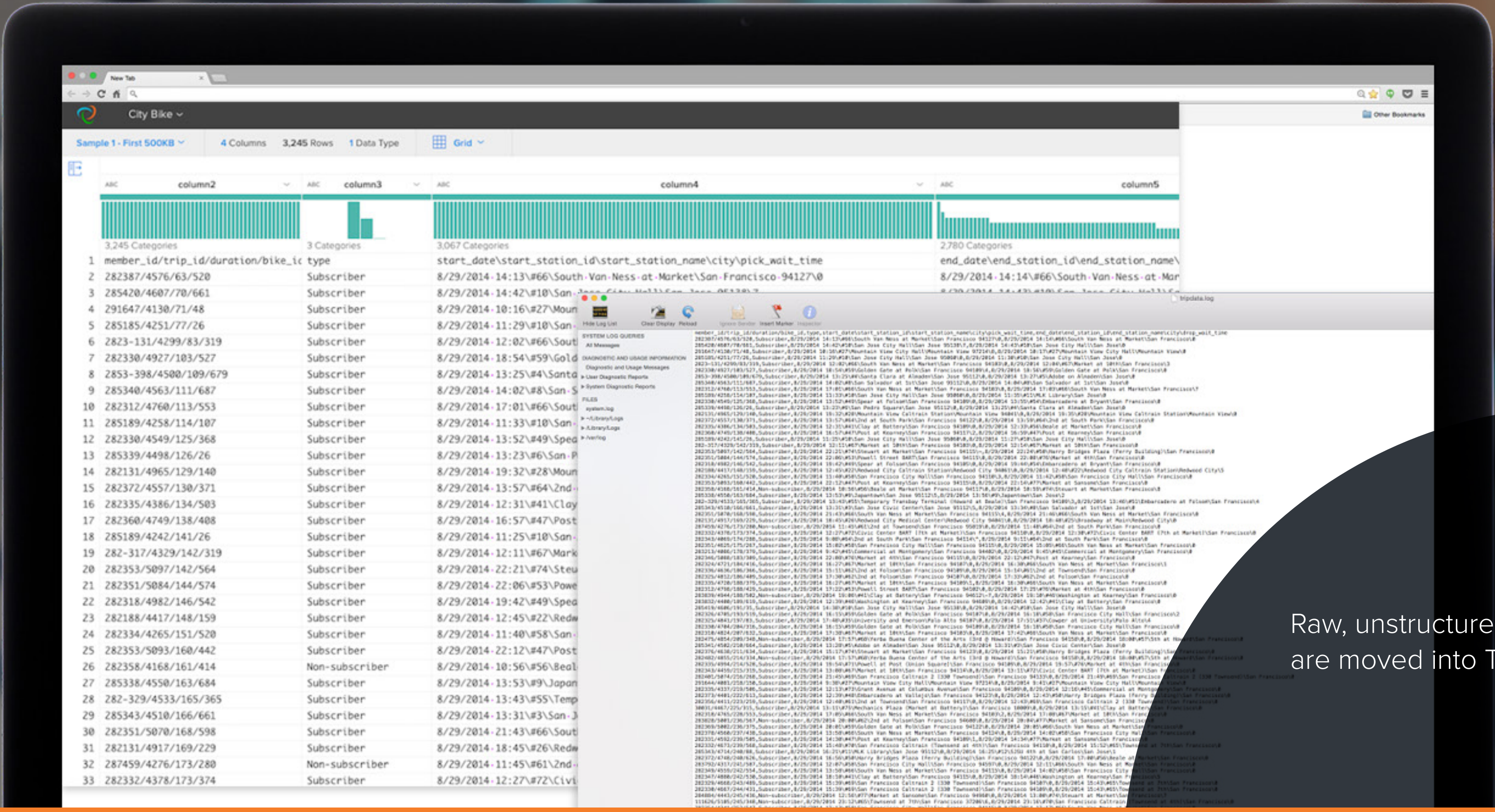
Customer tip: use the histograms to identify outlier values and to confirm your data follows the correct distribution.

Discovering

Trifacta's Interactive Exploration helps you discover features of your data and quickly determine the value of your dataset. Trifacta's data type inference, column-level profiles, interactive quality bars and histograms provide immediate visibility into trends and data issues, guiding the transformation process.

Discovering with Trifacta

The first dataset we'll be loading into Trifacta is the Trip data, as it requires the most structuring and cleaning. Bay Area Bike Share provides its data as a log file. Once loaded into Trifacta, you can immediately see that Trifacta automatically recommends some initial structuring of the dataset to help with analysis and to easily discover what is in your data.



	column2	column3	column4	column5
	3,245 Categories	3 Categories	3,067 Categories	2,780 Categories
1	member_id/trip_id/duration/bike_ic	type	start_date\start_station_id\start_station_name\city\pick_wait_time	end_date\end_station_id\end_station_name\
2	282387/4576/63/520	Subscriber	8/29/2014·14:13\#66\South·Van·Ness·at·Market\San·Francisco·94127\0	8/29/2014·14:14\#66\South·Van·Ness·at·Mar
3	285420/4607/70/661	Subscriber	8/29/2014·14:42\#10\San·	
4	291647/4130/71/48	Subscriber	8/29/2014·10:16\#27\Moun	
5	285185/4251/77/26	Subscriber	8/29/2014·11:29\#10\San·	
6	2823-131/4299/83/319	Subscriber	8/29/2014·12:02\#66\Sout	
7	282330/4927/103/527	Subscriber	8/29/2014·18:54\#59\Gold	
8	2853-398/4500/109/679	Subscriber	8/29/2014·13:25\#4\Santa	
9	285340/4563/111/687	Subscriber	8/29/2014·14:02\#8\San·S	
10	282312/4760/113/553	Subscriber	8/29/2014·17:01\#66\Sout	
11	285189/4258/114/107	Subscriber	8/29/2014·11:33\#10\San·	
12	282330/4549/125/368	Subscriber	8/29/2014·13:52\#49\Spea	
13	285339/4498/126/26	Subscriber	8/29/2014·13:23\#6\San·P	
14	282131/4965/129/140	Subscriber	8/29/2014·19:32\#28\Moun	
15	282372/4557/130/371	Subscriber	8/29/2014·13:57\#64\2nd·	
16	282335/4386/134/503	Subscriber	8/29/2014·12:31\#41\Clay	
17	282360/4749/138/408	Subscriber	8/29/2014·16:57\#47\Post	
18	285189/4242/141/26	Subscriber	8/29/2014·11:25\#10\San·	
19	282-317/4329/142/319	Subscriber	8/29/2014·12:11\#67\Mark	
20	282353/5097/142/564	Subscriber	8/29/2014·22:21\#74\Steu	
21	282351/5084/144/574	Subscriber	8/29/2014·22:06\#53\Powe	
22	282318/4982/146/542	Subscriber	8/29/2014·19:42\#49\Spea	
23	282188/4417/148/159	Subscriber	8/29/2014·12:45\#22\Redw	
24	282334/4265/151/520	Subscriber	8/29/2014·11:40\#58\San·	
25	282353/5093/160/442	Subscriber	8/29/2014·22:12\#47\Post	
26	282358/4168/161/414	Non-subscriber	8/29/2014·10:56\#56\Beal	
27	285338/4550/163/684	Subscriber	8/29/2014·13:53\#9\Japan	
28	282-329/4533/165/365	Subscriber	8/29/2014·13:43\#55\Temp	
29	285343/4510/166/661	Subscriber	8/29/2014·13:31\#3\San·J	
30	282351/5070/168/598	Subscriber	8/29/2014·21:43\#66\Sout	
31	282131/4917/169/229	Subscriber	8/29/2014·18:45\#26\Redw	
32	287459/4276/173/280	Non-subscriber	8/29/2014·11:45\#61\2nd·	
33	282332/4378/173/374	Subscriber	8/29/2014·12:27\#72\Civi	

Raw, unstructured log files are moved into Trifacta.

Customer tip: Trifacta's predictive transformation allows you to split a single data-rich column, such as one that stores addresses, into multiple columns. Analyzing street, city, state, and zip code separately can provide more meaningful geographic insights.

Structuring

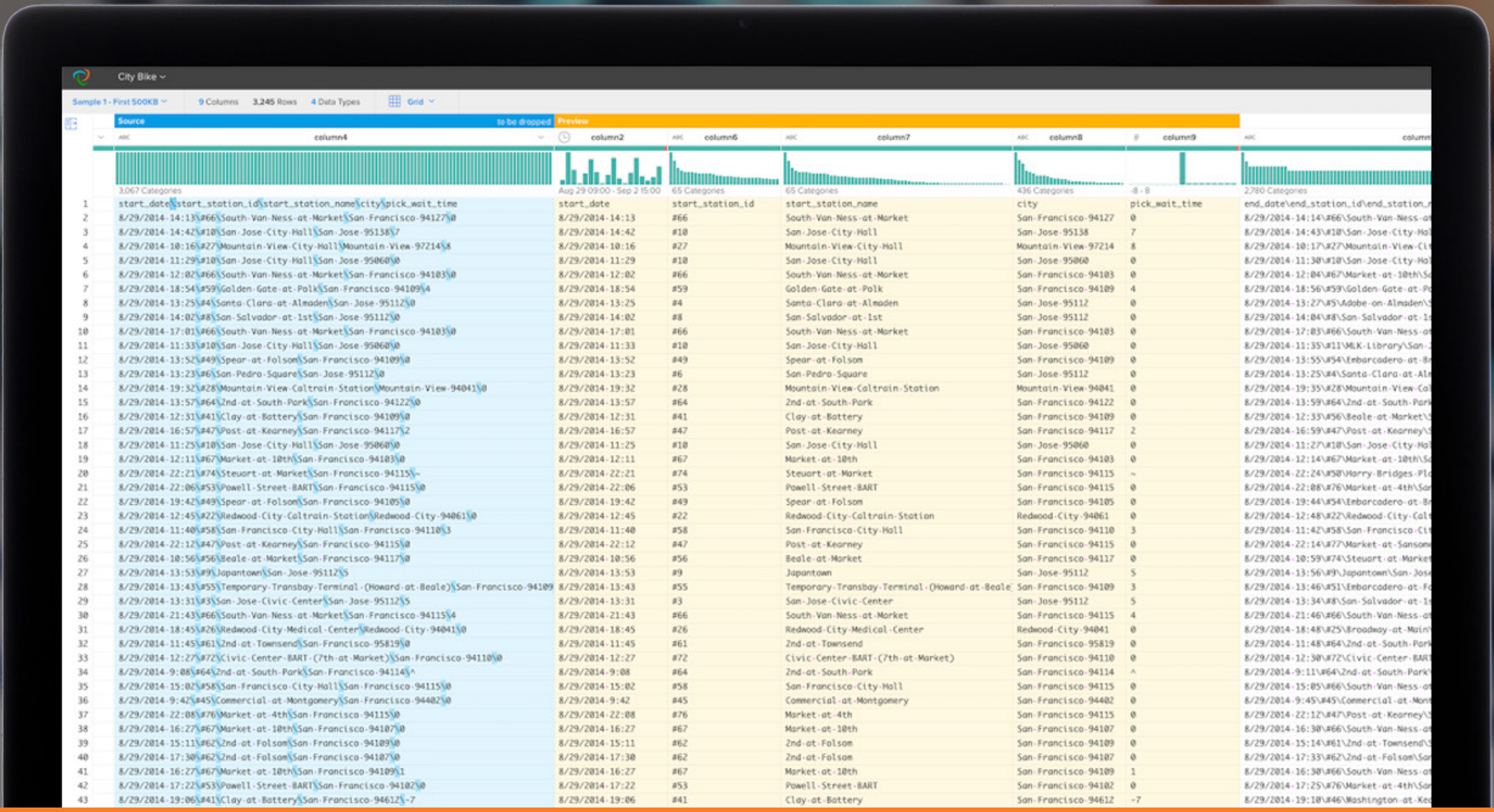
Structuring refers to actions that change the form or schema of your data. Splitting columns, pivoting rows and deleting fields are all forms of structuring.

Trifacta's Predictive Transformation allows you to simply highlight sections of your data to get suggestions of the appropriate transforms based on the data you're working with and the type of interaction you applied to the data.

Structuring with Trifacta

Although Trifacta Wrangler automatically recommends transforms to split the data into rows and columns, looking at the data there is still plenty of additional structuring to apply to this dataset. Data collected regarding date, station name and time for both the start and end period of each bike share session, are still lumped into a single column after this initial structuring provided by Trifacta. To give analysts the appropriate flexibility they will need to look into high traffic stations or peak borrow times, these attributes will need to split out into separate columns.

In the sample below, simply highlighting the backslash will generate a preview of column splits. To accept the split suggestion, hit "Add to Script" and the data will update and an additional step will be added to the script.



	column4	column2	column6	column7	column8	column9	column
	3,067 Categories	Aug 29 09:00 - Sep 2 15:00	65 Categories	65 Categories	436 Categories	-8 - 8	2,780 Categories
1	start_date\start_station_id\start_station_name\city\pick_wait_time	start_date	start_station_id	start_station_name	city	pick_wait_time	end_date\end_station_id\end_station_n
2	8/29/2014 14:13\#66\South Van Ness at Market\San Francisco 94127\0	8/29/2014 14:13	#66	South Van Ness at Market	San Francisco 94127	0	8/29/2014 14:14\#66\South Van Ness at
3	8/29/2014 14:42\#10\San Jose City Hall\San Jose 95138\7	8/29/2014 14:42	#10	San Jose City Hall	San Jose 95138	7	8/29/2014 14:43\#10\San Jose City Hal
4	8/29/2014 10:16\#27\Mountain View City Hall\Mountain View 97214\8	8/29/2014 10:16	#27	Mountain View City Hall	Mountain View 97214	8	8/29/2014 10:17\#27\Mountain View Cit
5	8/29/2014 11:29\#10\San Jose City Hall\San Jose 95060\0	8/29/2014 11:29	#10	San Jose City Hall	San Jose 95060	0	8/29/2014 11:30\#10\San Jose City Hal
6	8/29/2014 12:02\#66\South Van Ness at Market\San Francisco 94103\0	8/29/2014 12:02	#66	South Van Ness at Market	San Francisco 94103	0	8/29/2014 12:04\#67\Market at 10th\Sa
7	8/29/2014 18:54\#59\Golden Gate at Polk\San Francisco 94109\4	8/29/2014 18:54	#59	Golden Gate at Polk	San Francisco 94109	4	8/29/2014 18:56\#59\Golden Gate at P
8	8/29/2014 13:25\#4\Santa Clara at Almaden\San Jose 95112\0	8/29/2014 13:25	#4	Santa Clara at Almaden	San Jose 95112	0	8/29/2014 13:27\#5\Adobe on Almaden\S
9	8/29/2014 14:02\#8\San Salvador at 1st\San Jose 95112\0	8/29/2014 14:02	#8	San Salvador at 1st	San Jose 95112	0	8/29/2014 14:04\#8\San Salvador at 1s
10	8/29/2014 17:01\#66\South Van Ness at Market\San Francisco 94103\0	8/29/2014 17:01	#66	South Van Ness at Market	San Francisco 94103	0	8/29/2014 17:03\#66\South Van Ness at
11	8/29/2014 11:33\#10\San Jose City Hall\San Jose 95060\0	8/29/2014 11:33	#10	San Jose City Hall	San Jose 95060	0	8/29/2014 11:35\#11\MLK Library\San J
12	8/29/2014 13:52\#49\Spear at Folsom\San Francisco 94109\0	8/29/2014 13:52	#49	Spear at Folsom	San Francisco 94109	0	8/29/2014 13:55\#54\Embarcadero at B
13	8/29/2014 13:23\#6\San Pedro Square\San Jose 95112\0	8/29/2014 13:23	#6	San Pedro Square	San Jose 95112	0	8/29/2014 13:25\#4\Santa Clara at Alm
14	8/29/2014 19:32\#28\Mountain View Caltrain Station\Mountain View 94041\0	8/29/2014 19:32	#28	Mountain View Caltrain Station	Mountain View 94041	0	8/29/2014 19:35\#28\Mountain View Cal
15	8/29/2014 13:57\#64\2nd at South Park\San Francisco 94122\0	8/29/2014 13:57	#64	2nd at South Park	San Francisco 94122	0	8/29/2014 13:59\#64\2nd at South Park
16	8/29/2014 12:31\#41\Clay at Battery\San Francisco 94109\0	8/29/2014 12:31	#41	Clay at Battery	San Francisco 94109	0	8/29/2014 12:33\#56\Beale at Market\S
17	8/29/2014 16:57\#47\Post at Kearney\San Francisco 94117\2	8/29/2014 16:57	#47	Post at Kearney	San Francisco 94117	2	8/29/2014 16:59\#47\Post at Kearney\S
18	8/29/2014 11:25\#10\San Jose City Hall\San Jose 95060\0	8/29/2014 11:25	#10	San Jose City Hall	San Jose 95060	0	8/29/2014 11:27\#10\San Jose City Hal
19	8/29/2014 12:11\#67\Market at 10th\San Francisco 94103\0	8/29/2014 12:11	#67	Market at 10th	San Francisco 94103	0	8/29/2014 12:14\#67\Market at 10th\Sa
20	8/29/2014 22:21\#74\Steuart at Market\San Francisco 94115\~	8/29/2014 22:21	#74	Steuart at Market	San Francisco 94115	~	8/29/2014 22:24\#50\Harry Bridges Pl
21	8/29/2014 22:06\#53\Powell Street BART\San Francisco 94115\0	8/29/2014 22:06	#53	Powell Street BART	San Francisco 94115	0	8/29/2014 22:08\#76\Market at 4th\San
22	8/29/2014 19:42\#49\Spear at Folsom\San Francisco 94105\0	8/29/2014 19:42	#49	Spear at Folsom	San Francisco 94105	0	8/29/2014 19:44\#54\Embarcadero at B
23	8/29/2014 12:45\#22\Redwood City Caltrain Station\Redwood City 94061\0	8/29/2014 12:45	#22	Redwood City Caltrain Station	Redwood City 94061	0	8/29/2014 12:48\#22\Redwood City Calt
24	8/29/2014 11:40\#58\San Francisco City Hall\San Francisco 94110\3	8/29/2014 11:40	#58	San Francisco City Hall	San Francisco 94110	3	8/29/2014 11:42\#58\San Francisco Cit
25	8/29/2014 22:12\#47\Post at Kearney\San Francisco 94115\0	8/29/2014 22:12	#47	Post at Kearney	San Francisco 94115	0	8/29/2014 22:14\#77\Market at Sansom
26	8/29/2014 10:56\#56\Beale at Market\San Francisco 94117\0	8/29/2014 10:56	#56	Beale at Market	San Francisco 94117	0	8/29/2014 10:59\#74\Steuart at Market
27	8/29/2014 13:53\#9\Japantown\San Jose 95112\5	8/29/2014 13:53	#9	Japantown	San Jose 95112	5	8/29/2014 13:56\#9\Japantown\San Jose
28	8/29/2014 13:43\#55\Temporary Transbay Terminal (Howard at Beale)\San Francisco 94109	8/29/2014 13:43	#55	Temporary Transbay Terminal (Howard at Beale	San Francisco 94109	3	8/29/2014 13:46\#51\Embarcadero at F
29	8/29/2014 13:31\#3\San Jose Civic Center\San Jose 95112\5	8/29/2014 13:31	#3	San Jose Civic Center	San Jose 95112	5	8/29/2014 13:34\#8\San Salvador at 1s
30	8/29/2014 21:43\#66\South Van Ness at Market\San Francisco 94115\4	8/29/2014 21:43	#66	South Van Ness at Market	San Francisco 94115	4	8/29/2014 21:46\#66\South Van Ness at
31	8/29/2014 18:45\#26\Redwood City Medical Center\Redwood City 94041\0	8/29/2014 18:45	#26	Redwood City Medical Center	Redwood City 94041	0	8/29/2014 18:48\#25\Broadway at Main\
32	8/29/2014 11:45\#61\2nd at Townsend\San Francisco 95819\0	8/29/2014 11:45	#61	2nd at Townsend	San Francisco 95819	0	8/29/2014 11:48\#64\2nd at South Park
33	8/29/2014 12:27\#72\Civic Center BART (7th at Market)\San Francisco 94110\0	8/29/2014 12:27	#72	Civic Center BART (7th at Market)	San Francisco 94110	0	8/29/2014 12:30\#72\Civic Center BART
34	8/29/2014 9:08\#64\2nd at South Park\San Francisco 94114\^	8/29/2014 9:08	#64	2nd at South Park	San Francisco 94114	^	8/29/2014 9:11\#64\2nd at South Park\
35	8/29/2014 15:02\#58\San Francisco City Hall\San Francisco 94115\0	8/29/2014 15:02	#58	San Francisco City Hall	San Francisco 94115	0	8/29/2014 15:05\#66\South Van Ness at
36	8/29/2014 9:42\#45\Commercial at Montgomery\San Francisco 94402\0	8/29/2014 9:42	#45	Commercial at Montgomery	San Francisco 94402	0	8/29/2014 9:45\#45\Commercial at Mont
37	8/29/2014 22:08\#76\Market at 4th\San Francisco 94115\0	8/29/2014 22:08	#76	Market at 4th	San Francisco 94115	0	8/29/2014 22:12\#47\Post at Kearney\S
38	8/29/2014 16:27\#67\Market at 10th\San Francisco 94107\0	8/29/2014 16:27	#67	Market at 10th	San Francisco 94107	0	8/29/2014 16:30\#66\South Van Ness at
39	8/29/2014 15:11\#62\2nd at Folsom\San Francisco 94109\0	8/29/2014 15:11	#62	2nd at Folsom	San Francisco 94109	0	8/29/2014 15:14\#61\2nd at Townsend\S
40	8/29/2014 17:30\#62\2nd at Folsom\San Francisco 94107\0	8/29/2014 17:30	#62	2nd at Folsom	San Francisco 94107	0	8/29/2014 17:33\#62\2nd at Folsom\San
41	8/29/2014 16:27\#67\Market at 10th\San Francisco 94109\1	8/29/2014 16:27	#67	Market at 10th	San Francisco 94109	1	8/29/2014 16:30\#66\South Van Ness at
42	8/29/2014 17:22\#53\Powell Street BART\San Francisco 94102\0	8/29/2014 17:22	#53	Powell Street BART	San Francisco 94102	0	8/29/2014 17:25\#76\Market at 4th\San
43	8/29/2014 19:06\#41\Clay at Battery\San Francisco 94612\-7	8/29/2014 19:06	#41	Clay at Battery	San Francisco 94612	-7	8/29/2014 19:10\#46\Washington at Ke

Customer tip: Select the red section of the data quality bar to isolate invalid data and delete or correct the bad values in a single step.

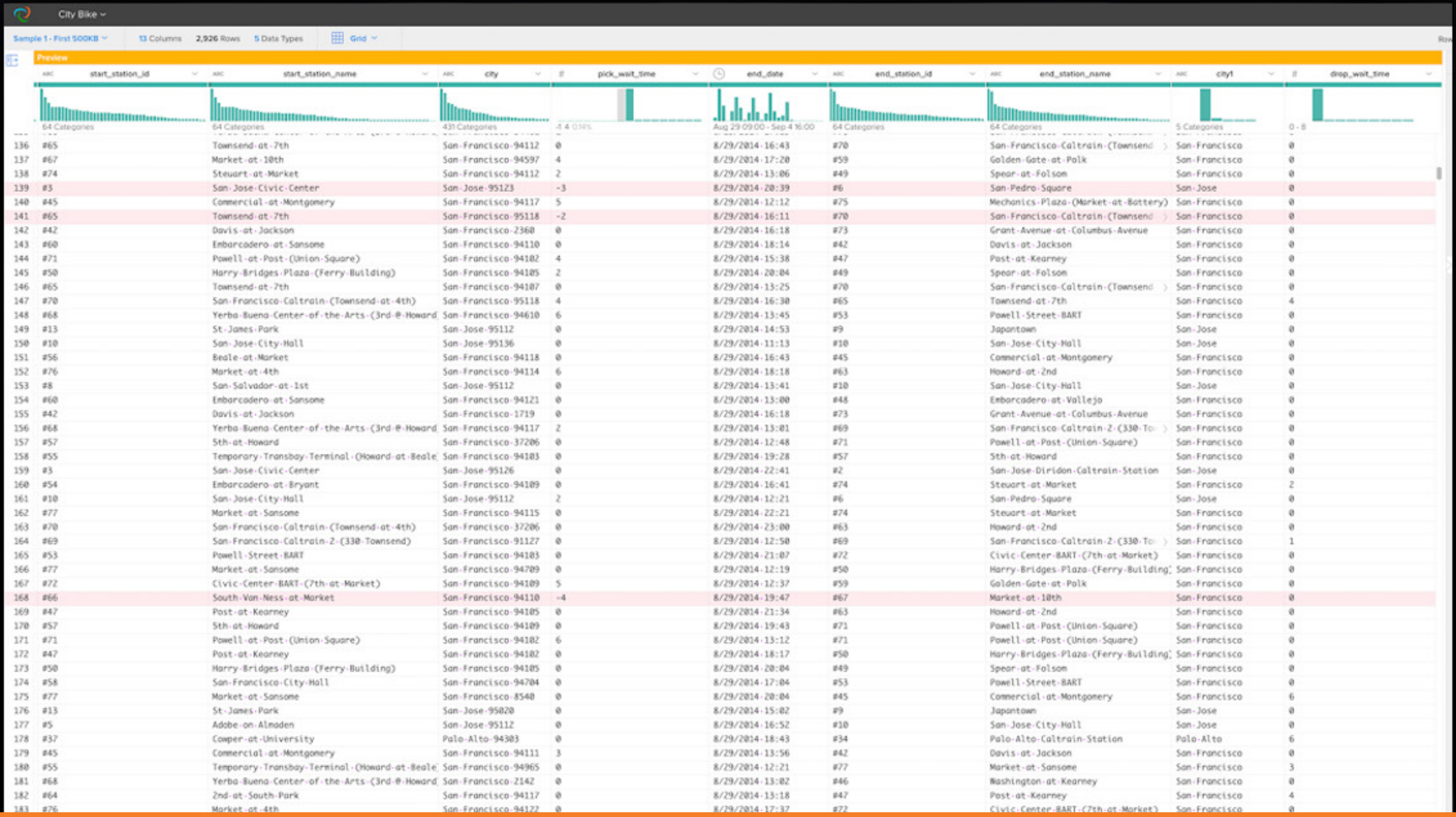
Cleaning

During the cleaning stage, users identify data quality issues, such as missing or mismatched values, and apply the appropriate transformation to correct or delete these values from the dataset. In Trifacta, you can isolate and replace null values (which would otherwise throw off your analysis) with a single click.

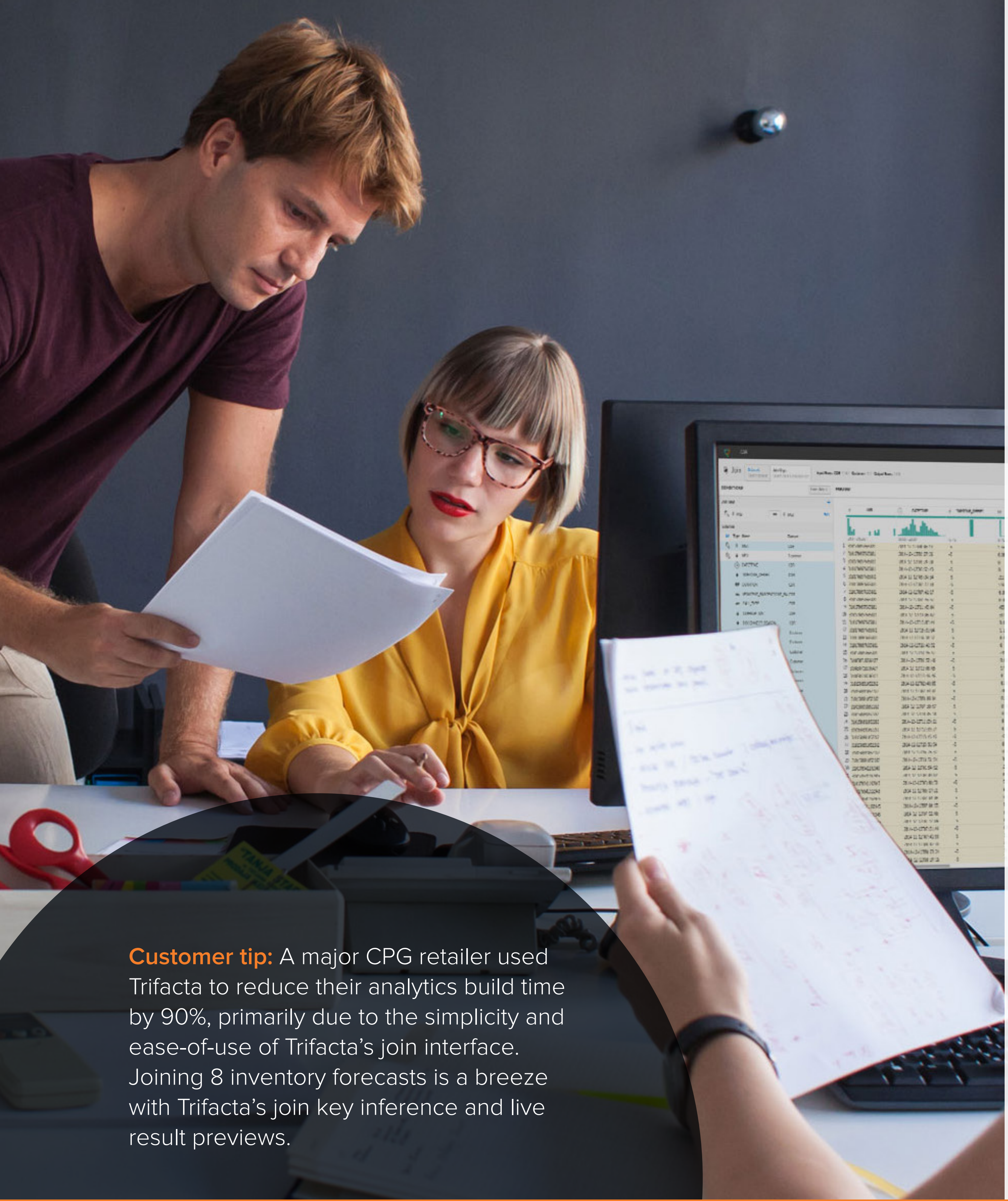
Cleaning with Trifacta

Once you've structured the data and dropped unwanted columns, you can further clean the missing, mismatched or otherwise unnecessary data that may affect your analysis.

In this dataset, the Bay Area Bike Share program also collects negative wait times, which you can identify in the histogram. In reviewing the bike wait time column, we can see that there are some negative values present in the column. Given that it's impossible for a biker to wait a negative time for a bike, we will simply remove those rows since they seem to represent errors in the dataset and might throw off our analysis. The histogram shows the range of values in the dataset. By brushing over the portion of values to the left of zero, you can highlight all of the negative values in that column and delete them.



	start_station_id	start_station_name	city	pick_wait_time	end_date	end_station_id	end_station_name	city1	drop_wait_time
	64 Categories	64 Categories	431 Categories	-4 4 0.14%	Aug 29 09:00 - Sep 4 16:00	64 Categories	64 Categories	5 Categories	0 - 8
136	#65	Townsend at 7th	San Francisco 94112	0	8/29/2014 16:43	#70	San Francisco Caltrain (Townsend	San Francisco	0
137	#67	Market at 10th	San Francisco 94597	4	8/29/2014 17:20	#59	Golden Gate at Polk	San Francisco	0
138	#74	Steuart at Market	San Francisco 94112	2	8/29/2014 13:06	#49	Spear at Folsom	San Francisco	0
139	#3	San Jose Civic Center	San Jose 95123	-3	8/29/2014 20:39	#6	San Pedro Square	San Jose	0
140	#45	Commercial at Montgomery	San Francisco 94117	5	8/29/2014 12:12	#75	Mechanics Plaza (Market at Battery)	San Francisco	0
141	#65	Townsend at 7th	San Francisco 95118	-2	8/29/2014 16:11	#70	San Francisco Caltrain (Townsend	San Francisco	0
142	#42	Davis at Jackson	San Francisco 2360	0	8/29/2014 16:18	#73	Grant Avenue at Columbus Avenue	San Francisco	0
143	#60	Embarcadero at Sansome	San Francisco 94110	0	8/29/2014 18:14	#42	Davis at Jackson	San Francisco	0
144	#71	Powell at Post (Union Square)	San Francisco 94102	4	8/29/2014 15:38	#47	Post at Kearney	San Francisco	0
145	#50	Harry Bridges Plaza (Ferry Building)	San Francisco 94105	2	8/29/2014 20:04	#49	Spear at Folsom	San Francisco	0
146	#65	Townsend at 7th	San Francisco 94107	0	8/29/2014 13:25	#70	San Francisco Caltrain (Townsend	San Francisco	0
147	#70	San Francisco Caltrain (Townsend at 4th)	San Francisco 95118	4	8/29/2014 16:30	#65	Townsend at 7th	San Francisco	4
148	#68	Yerba Buena Center of the Arts (3rd @ Howard	San Francisco 94610	6	8/29/2014 13:45	#53	Powell Street BART	San Francisco	0
149	#13	St James Park	San Jose 95112	0	8/29/2014 14:53	#9	Japantown	San Jose	0
150	#10	San Jose City Hall	San Jose 95136	0	8/29/2014 11:13	#10	San Jose City Hall	San Jose	0
151	#56	Beale at Market	San Francisco 94118	0	8/29/2014 16:43	#45	Commercial at Montgomery	San Francisco	0
152	#76	Market at 4th	San Francisco 94114	6	8/29/2014 18:18	#63	Howard at 2nd	San Francisco	0
153	#8	San Salvador at 1st	San Jose 95112	0	8/29/2014 13:41	#10	San Jose City Hall	San Jose	0
154	#60	Embarcadero at Sansome	San Francisco 94121	0	8/29/2014 13:00	#48	Embarcadero at Vallejo	San Francisco	0
155	#42	Davis at Jackson	San Francisco 1719	0	8/29/2014 16:18	#73	Grant Avenue at Columbus Avenue	San Francisco	0
156	#68	Yerba Buena Center of the Arts (3rd @ Howard	San Francisco 94117	2	8/29/2014 13:01	#69	San Francisco Caltrain 2 (330 To	San Francisco	0
157	#57	5th at Howard	San Francisco 37206	0	8/29/2014 12:48	#71	Powell at Post (Union Square)	San Francisco	0
158	#55	Temporary Transbay Terminal (Howard at Beale	San Francisco 94103	0	8/29/2014 19:28	#57	5th at Howard	San Francisco	0
159	#3	San Jose Civic Center	San Jose 95126	0	8/29/2014 22:41	#2	San Jose Diridon Caltrain Station	San Jose	0
160	#54	Embarcadero at Bryant	San Francisco 94109	0	8/29/2014 16:41	#74	Steuart at Market	San Francisco	2
161	#10	San Jose City Hall	San Jose 95112	2	8/29/2014 12:21	#6	San Pedro Square	San Jose	0
162	#77	Market at Sansome	San Francisco 94115	0	8/29/2014 22:21	#74	Steuart at Market	San Francisco	0
163	#70	San Francisco Caltrain (Townsend at 4th)	San Francisco 37206	0	8/29/2014 23:00	#63	Howard at 2nd	San Francisco	0
164	#69	San Francisco Caltrain 2 (330 Townsend)	San Francisco 91127	0	8/29/2014 12:50	#69	San Francisco Caltrain 2 (330 To	San Francisco	1
165	#53	Powell Street BART	San Francisco 94103	0	8/29/2014 21:07	#72	Civic Center BART (7th at Market)	San Francisco	0
166	#77	Market at Sansome	San Francisco 94709	0	8/29/2014 12:19	#50	Harry Bridges Plaza (Ferry Building	San Francisco	0
167	#72	Civic Center BART (7th at Market)	San Francisco 94109	5	8/29/2014 12:37	#59	Golden Gate at Polk	San Francisco	0
168	#66	South Van Ness at Market	San Francisco 94110	-4	8/29/2014 19:47	#67	Market at 10th	San Francisco	0
169	#47	Post at Kearney	San Francisco 94105	0	8/29/2014 21:34	#63	Howard at 2nd	San Francisco	0
170	#57	5th at Howard	San Francisco 94109	0	8/29/2014 19:43	#71	Powell at Post (Union Square)	San Francisco	0
171	#71	Powell at Post (Union Square)	San Francisco 94102	6	8/29/2014 13:12	#71	Powell at Post (Union Square)	San Francisco	0
172	#47	Post at Kearney	San Francisco 94102	0	8/29/2014 18:17	#50	Harry Bridges Plaza (Ferry Building	San Francisco	0
173	#50	Harry Bridges Plaza (Ferry Building)	San Francisco 94105	0	8/29/2014 20:04	#49	Spear at Folsom	San Francisco	0
174	#58	San Francisco City Hall	San Francisco 94704	0	8/29/2014 17:04	#53	Powell Street BART	San Francisco	0
175	#77	Market at Sansome	San Francisco 8540	0	8/29/2014 20:04	#45	Commercial at Montgomery	San Francisco	6
176	#13	St James Park	San Jose 95020	0	8/29/2014 15:02	#9	Japantown	San Jose	0
177	#5	Adobe on Almaden	San Jose 95112	0	8/29/2014 16:52	#10	San Jose City Hall	San Jose	0
178	#37	Cowper at University	Palo Alto 94303	0	8/29/2014 18:43	#34	Palo Alto Caltrain Station	Palo Alto	6
179	#45	Commercial at Montgomery	San Francisco 94111	3	8/29/2014 13:56	#42	Davis at Jackson	San Francisco	0
180	#55	Temporary Transbay Terminal (Howard at Beale	San Francisco 94965	0	8/29/2014 12:21	#77	Market at Sansome	San Francisco	3
181	#68	Yerba Buena Center of the Arts (3rd @ Howard	San Francisco 2142	0	8/29/2014 13:02	#46	Washington at Kearney	San Francisco	0
182	#64	2nd at South Park	San Francisco 94117	0	8/29/2014 13:18	#47	Post at Kearney	San Francisco	4
183	#76	Market at 4th	San Francisco 94122	0	8/29/2014 17:37	#72	Civic Center BART (7th at Market)	San Francisco	0



Customer tip: A major CPG retailer used Trifacta to reduce their analytics build time by 90%, primarily due to the simplicity and ease-of-use of Trifacta's join interface. Joining 8 inventory forecasts is a breeze with Trifacta's join key inference and live result previews.

Enriching

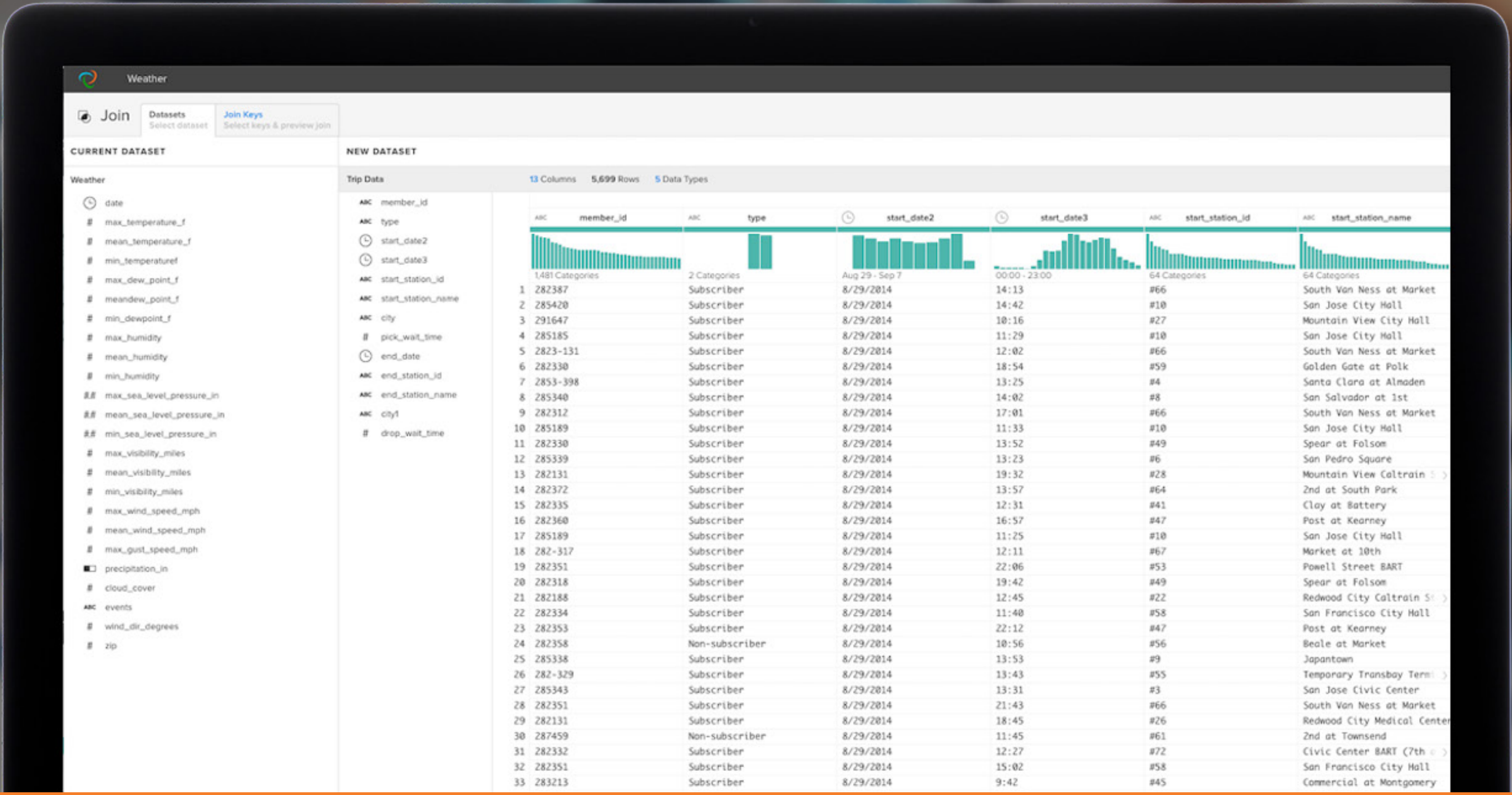
The data needed to make business decisions can often be spread across multiple files. In order to gather all the necessary insights, you need to enrich your existing dataset by joining and aggregating multiple data sources.

Trifacta's data enrichment features allow you to easily execute lookups to data dictionaries or execute joins with disparate datasets. Rather than keying in a command, Trifacta's intelligent join inference uses machine learning to rapidly identify appropriate join keys across your diverse datasets.

Enriching with Trifacta

Now that the Trip data is structured, cleaned and standardized, you can start enriching the data. As mentioned before, weather can be an important factor in bike usage, as it is much less appealing to ride a bicycle on a particularly rainy or windy day. An analyst might be able to get further insights and draw more well-informed conclusions if you added weather data to the Trip dataset. To accomplish this, you can use the join function to bring these two datasets together.

After loading the Weather data you can see that it is already quite clean and standardized. Next, hitting the join function will allow you to select the Trip dataset that you've already cleaned. You want to join the two datasets using the date value as your join key so that the rows and columns match--however only one date column needs to be incorporated into the output dataset to avoid a duplicate. Once the join is created, this transformed dataset will have both trip and weather data within it.



Weather

- date
- max_temperature_f
- mean_temperature_f
- min_temperaturef
- max_dew_point_f
- meandew_point_f
- min_dewpoint_f
- max_humidity
- mean_humidity
- min_humidity
- max_sea_level_pressure_in
- mean_sea_level_pressure_in
- min_sea_level_pressure_in
- max_visibility_miles
- mean_visibility_miles
- min_visibility_miles
- max_wind_speed_mph
- mean_wind_speed_mph
- max_gust_speed_mph
- precipitation_in
- cloud_cover
- events
- wind_dir_degrees
- zip

Trip Data

- member_id
- type
- start_date2
- start_date3
- start_station_id
- start_station_name
- city
- pick_wait_time
- end_date
- end_station_id
- end_station_name
- city1
- drop_wait_time

13 Columns 5,699 Rows 5 Data Types

	member_id	type	start_date2	start_date3	start_station_id	start_station_name
	1,481 Categories	2 Categories	Aug 29 - Sep 7	00:00 - 23:00	64 Categories	64 Categories
1	282387	Subscriber	8/29/2014	14:13	#66	South Van Ness at Market
2	285420	Subscriber	8/29/2014	14:42	#10	San Jose City Hall
3	291647	Subscriber	8/29/2014	10:16	#27	Mountain View City Hall
4	285185	Subscriber	8/29/2014	11:29	#10	San Jose City Hall
5	2823-131	Subscriber	8/29/2014	12:02	#66	South Van Ness at Market
6	282330	Subscriber	8/29/2014	18:54	#59	Golden Gate at Polk
7	2853-398	Subscriber	8/29/2014	13:25	#4	Santa Clara at Almaden
8	285340	Subscriber	8/29/2014	14:02	#8	San Salvador at 1st
9	282312	Subscriber	8/29/2014	17:01	#66	South Van Ness at Market
10	285189	Subscriber	8/29/2014	11:33	#10	San Jose City Hall
11	282330	Subscriber	8/29/2014	13:52	#49	Spear at Folsom
12	285339	Subscriber	8/29/2014	13:23	#6	San Pedro Square
13	282131	Subscriber	8/29/2014	19:32	#28	Mountain View Caltrain S
14	282372	Subscriber	8/29/2014	13:57	#64	2nd at South Park
15	282335	Subscriber	8/29/2014	12:31	#41	Clay at Battery
16	282360	Subscriber	8/29/2014	16:57	#47	Post at Kearney
17	285189	Subscriber	8/29/2014	11:25	#10	San Jose City Hall
18	282-317	Subscriber	8/29/2014	12:11	#67	Market at 10th
19	282351	Subscriber	8/29/2014	22:06	#53	Powell Street BART
20	282318	Subscriber	8/29/2014	19:42	#49	Spear at Folsom
21	282188	Subscriber	8/29/2014	12:45	#22	Redwood City Caltrain St
22	282334	Subscriber	8/29/2014	11:40	#58	San Francisco City Hall
23	282353	Subscriber	8/29/2014	22:12	#47	Post at Kearney
24	282358	Non-subscriber	8/29/2014	10:56	#56	Beale at Market
25	285338	Subscriber	8/29/2014	13:53	#9	Japantown
26	282-329	Subscriber	8/29/2014	13:43	#55	Temporary Transbay Term
27	285343	Subscriber	8/29/2014	13:31	#3	San Jose Civic Center
28	282351	Subscriber	8/29/2014	21:43	#66	South Van Ness at Market
29	282131	Subscriber	8/29/2014	18:45	#26	Redwood City Medical Center
30	287459	Non-subscriber	8/29/2014	11:45	#61	2nd at Townsend
31	282332	Subscriber	8/29/2014	12:27	#72	Civic Center BART (7th
32	282351	Subscriber	8/29/2014	15:02	#58	San Francisco City Hall
33	283213	Subscriber	8/29/2014	9:42	#45	Commercial at Montgomery

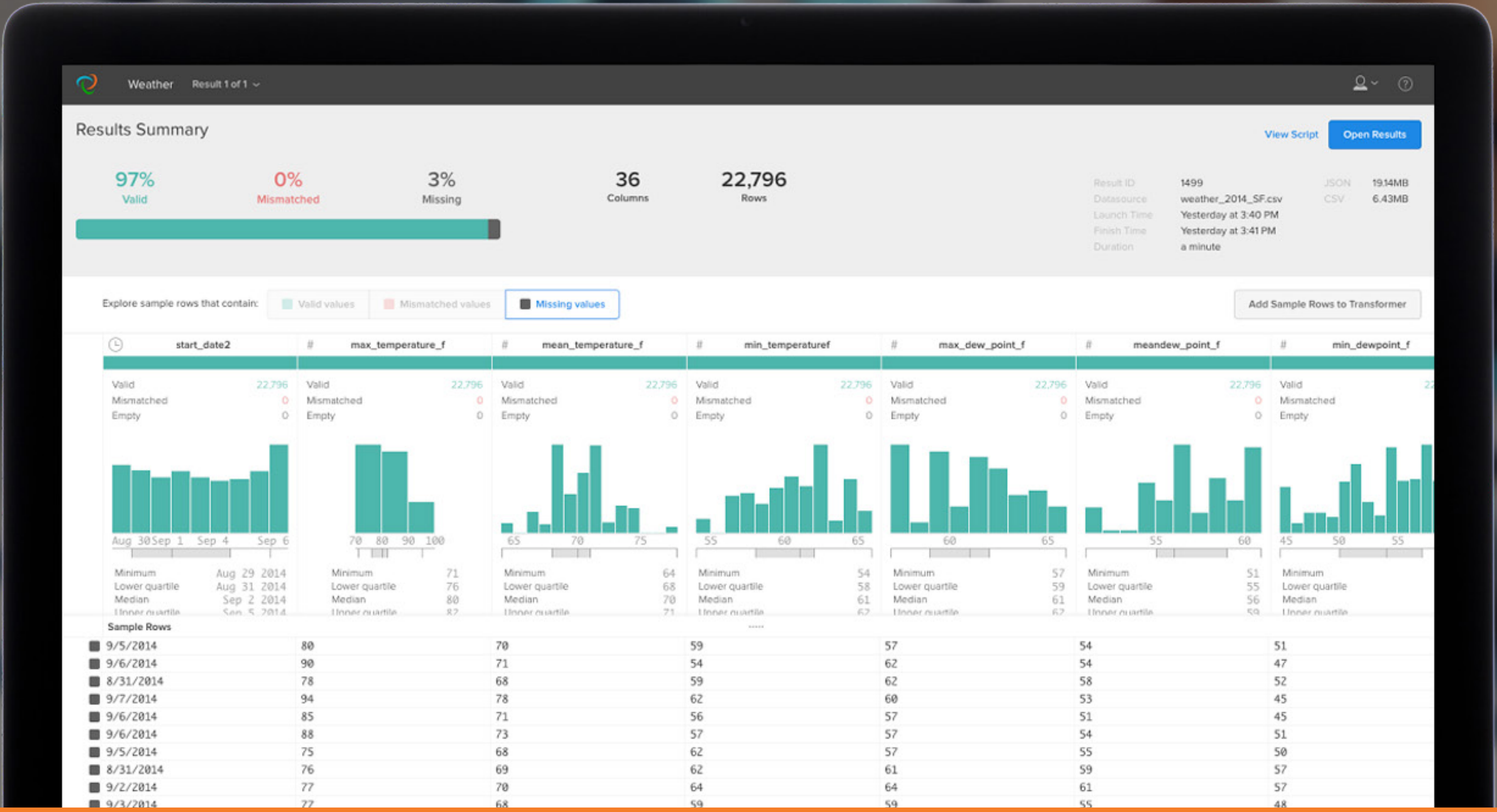
Customer tip: Data wrangling is an iterative process, so Trifacta allows you to review your result dataset for further refining based on direction from the job results window.

Validating

Trifacta displays the final output of your transformation output in the Job Results view across the complete dataset. Here, you can do a final check for any missing or mismatched data that wasn't corrected in the transformation process. After your data has been initially wrangled, you need to validate that your output dataset has the intended structure and content before publishing for broader analysis.

Validating with Trifacta

Once the transformations are complete, you can view the Results Summary, which displays detailed statistics of the transformations applied over the entire dataset. You can then export the results of this transformation into the appropriate output format best-fit for your visualization or analytics tool of choice, such as Tableau.



Results Summary

97% Valid

0% Mismatched

3% Missing

36 Columns

22,796 Rows

Result ID	1499
Datasource	weather_2014_SF.csv
Launch Time	Yesterday at 3:40 PM
Finish Time	Yesterday at 3:41 PM
Duration	a minute

JSON	19.14MB
CSV	6.43MB

Explore sample rows that contain:

	start_date2	max_temperature_f	mean_temperature_f	min_temperaturef	max_dew_point_f	meandew_point_f	min_dewpoint_f
Valid	22,796	22,796	22,796	22,796	22,796	22,796	22
Mismatched	0	0	0	0	0	0	
Empty	0	0	0	0	0	0	
Minimum	Aug 29 2014	71	64	54	57	51	
Lower quartile	Aug 31 2014	76	68	58	59	55	
Median	Sep 2 2014	80	70	61	61	56	
Upper quartile	Sep 5 2014	82	71	62	62	59	

Sample Rows

start_date2	max_temperature_f	mean_temperature_f	min_temperaturef	max_dew_point_f	meandew_point_f	min_dewpoint_f
9/5/2014	80	70	59	57	54	51
9/6/2014	90	71	54	62	54	47
8/31/2014	78	68	59	62	58	52
9/7/2014	94	78	62	60	53	45
9/6/2014	85	71	56	57	51	45
9/6/2014	88	73	57	57	54	51
9/5/2014	75	68	62	57	55	50
8/31/2014	76	69	62	61	59	57
9/2/2014	77	70	64	64	61	57
9/3/2014	77	68	59	59	55	48

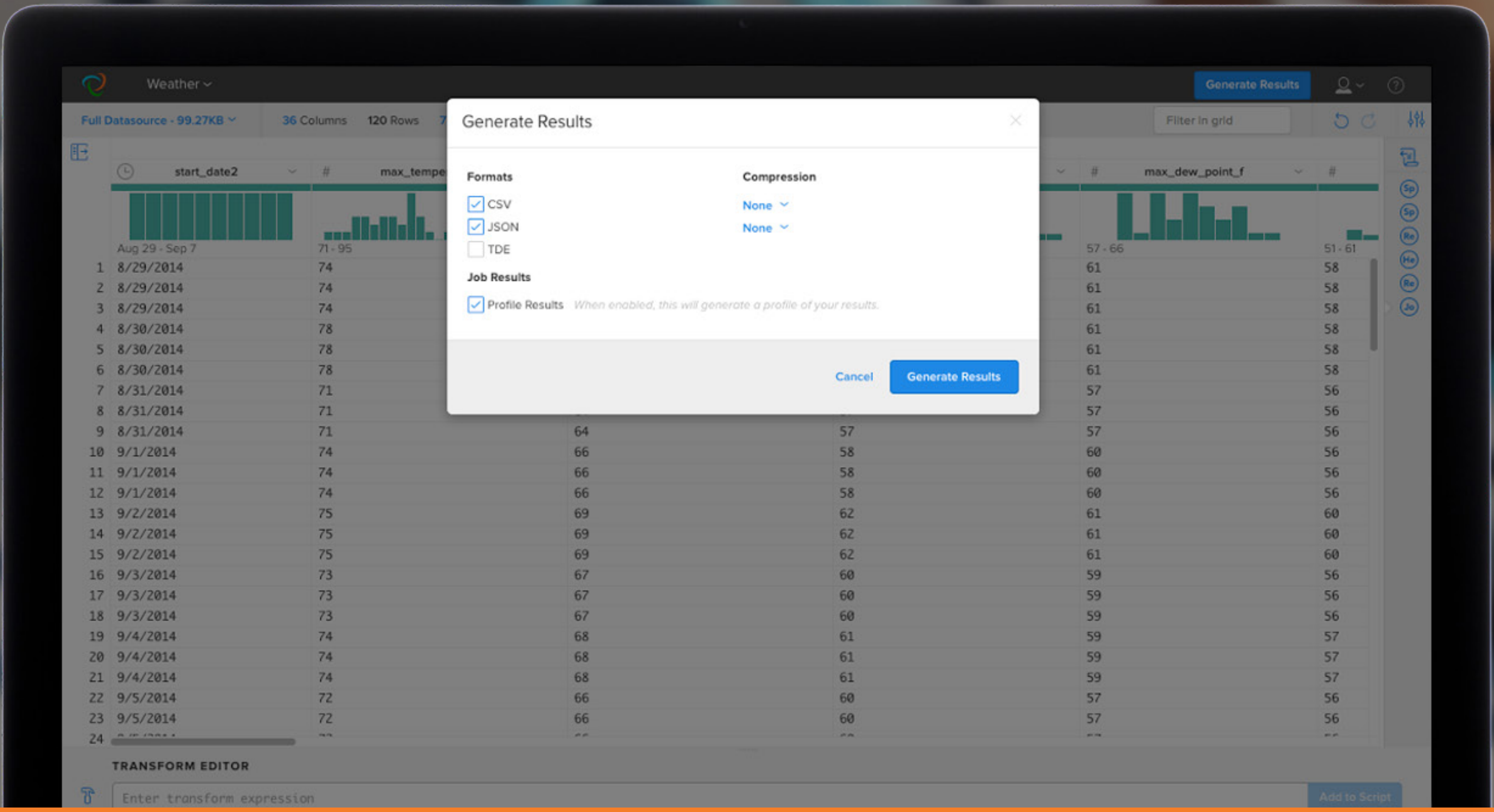
Publishing

When your data has been successfully structured, cleaned, enriched and validated, it's time to publish your wrangled output for use in downstream analytics processes.

Through the wrangling process, a wider variety of data sources can be used in different statistics, analytics and data visualization applications. This broadens the usage of data throughout the organization and enhances the potential value of data to the business.

Publishing with Trifacta

Now you're ready to deliver the output of your data wrangling efforts into the appropriate format for downstream analytic uses. You can publish and save your results as a CSV, JSON, or Tableau Data Extract (TDE) files. The dataset will remain in the Trifacta Workspace in case you need to make additional transformations later on.



Generate Results

Formats

☑ CSV
☑ JSON
□ TDE

Compression

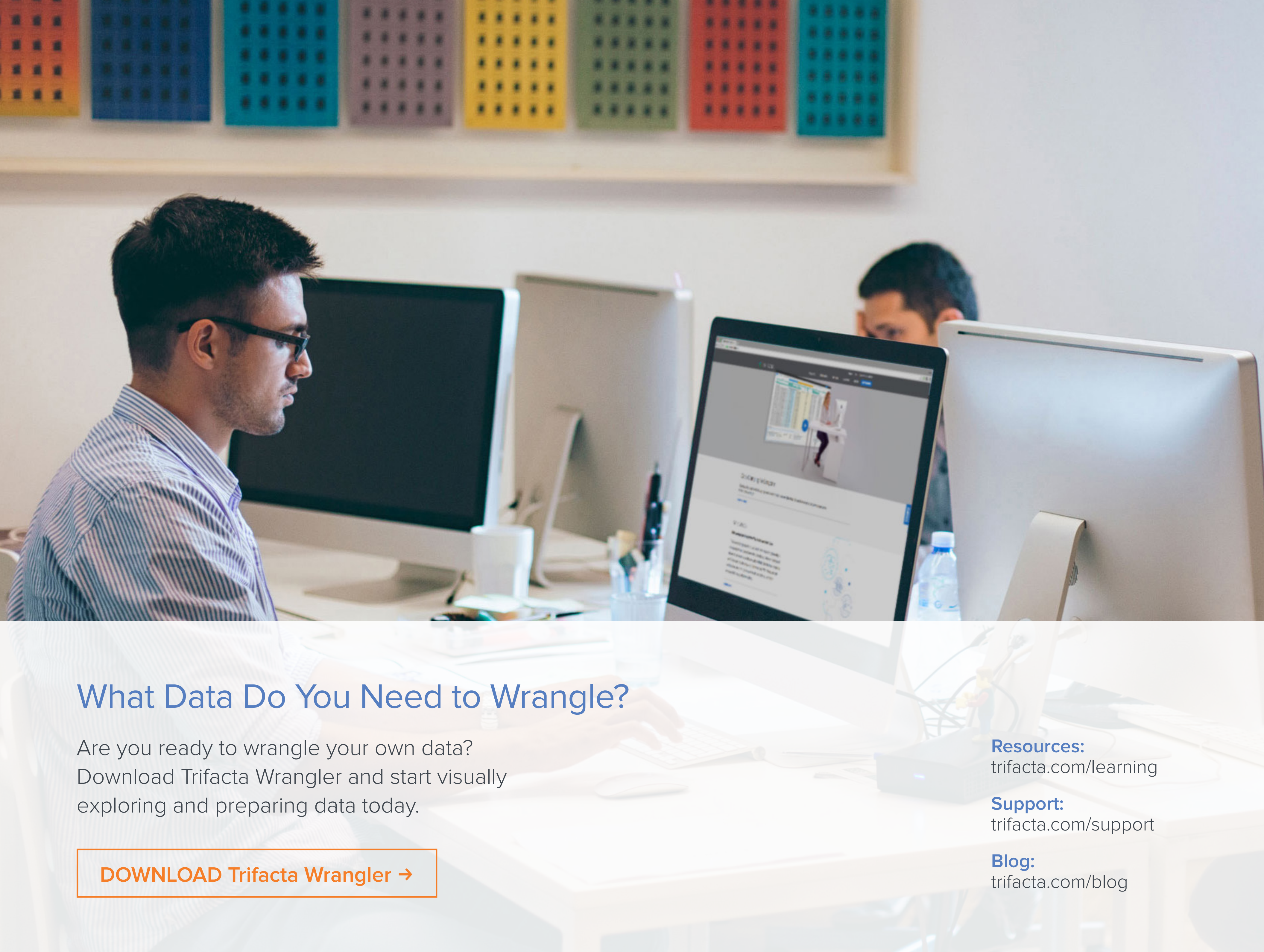
CSV: None
JSON: None

Job Results

☑ Profile Results *When enabled, this will generate a profile of your results.*

Cancel | Generate Results

	start_date2	max_tempe			max_dew_point_f	#
	Aug 29 - Sep 7	71 - 95			57 - 66	51 - 61
1	8/29/2014	74			61	58
2	8/29/2014	74			61	58
3	8/29/2014	74			61	58
4	8/30/2014	78			61	58
5	8/30/2014	78			61	58
6	8/30/2014	78			61	58
7	8/31/2014	71			57	56
8	8/31/2014	71			57	56
9	8/31/2014	71	64	57	57	56
10	9/1/2014	74	66	58	60	56
11	9/1/2014	74	66	58	60	56
12	9/1/2014	74	66	58	60	56
13	9/2/2014	75	69	62	61	60
14	9/2/2014	75	69	62	61	60
15	9/2/2014	75	69	62	61	60
16	9/3/2014	73	67	60	59	56
17	9/3/2014	73	67	60	59	56
18	9/3/2014	73	67	60	59	56
19	9/4/2014	74	68	61	59	57
20	9/4/2014	74	68	61	59	57
21	9/4/2014	74	68	61	59	57
22	9/5/2014	72	66	60	57	56
23	9/5/2014	72	66	60	57	56
24	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]



What Data Do You Need to Wrangle?

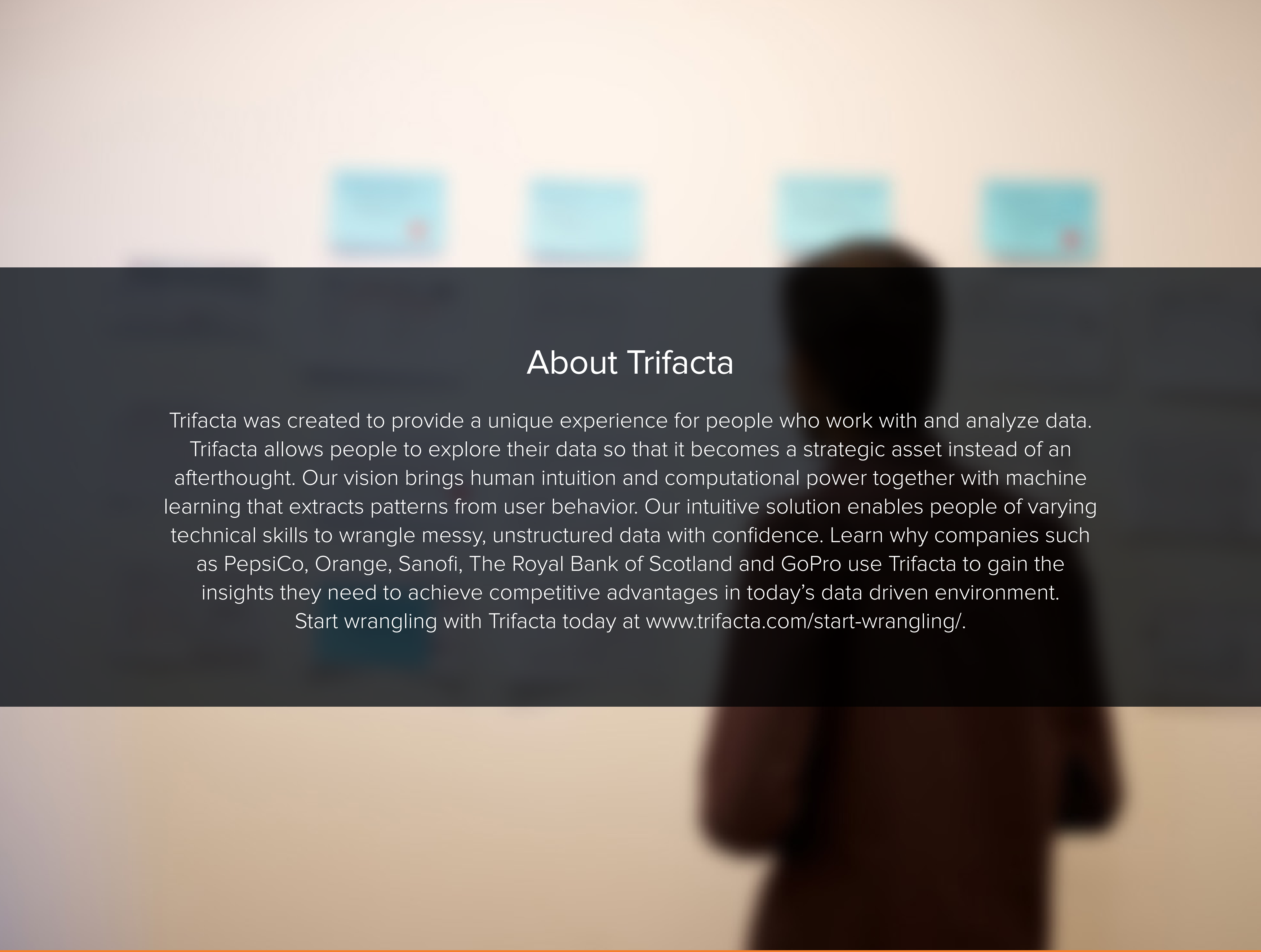
Are you ready to wrangle your own data? Download Trifacta Wrangler and start visually exploring and preparing data today.

DOWNLOAD Trifacta Wrangler →

Resources:
trifacta.com/learning

Support:
trifacta.com/support

Blog:
trifacta.com/blog



About Trifacta

Trifacta was created to provide a unique experience for people who work with and analyze data. Trifacta allows people to explore their data so that it becomes a strategic asset instead of an afterthought. Our vision brings human intuition and computational power together with machine learning that extracts patterns from user behavior. Our intuitive solution enables people of varying technical skills to wrangle messy, unstructured data with confidence. Learn why companies such as PepsiCo, Orange, Sanofi, The Royal Bank of Scotland and GoPro use Trifacta to gain the insights they need to achieve competitive advantages in today’s data driven environment.
Start wrangling with Trifacta today at www.trifacta.com/start-wrangling/.

TRIFACTA