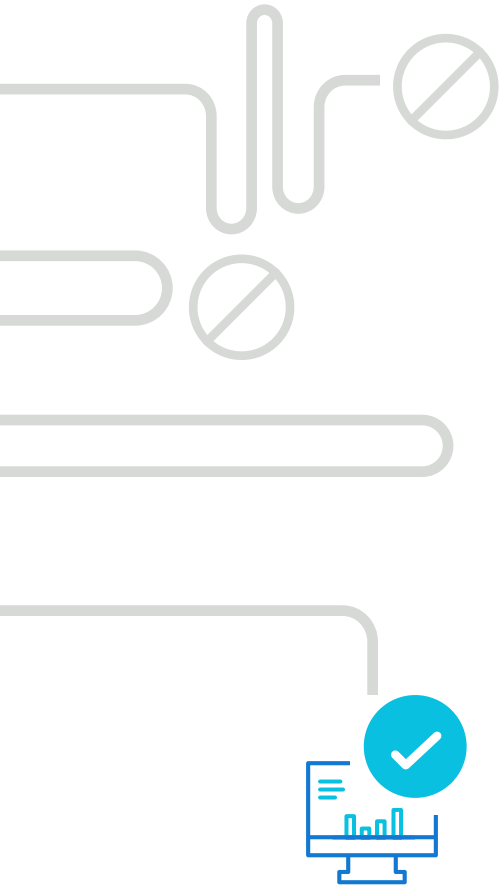
Self-Service Data Preparation: The "Goldilocks" Solution to Integrating Unfamiliar Client Data

TRIFACTA®

When Excel is too limited and code is too technical, self-service data preparation is just right.



Let's say you work for an analytics provider—whether marketing, healthcare, supply chain or other analytics—and you've just closed a deal with a new client. One of your first orders of business is to get your hands on your client's data so you can start delivering value as quickly as possible.

Your client sets up data sharing. Your team stands ready to onboard your client's data. The process involves connecting to, understanding, cleaning, blending, and outputting the client data into a workflow that, in turn, feeds your analytics product or platform.

Integrating unfamiliar client data involves connecting to, understanding, cleaning, blending and outputting client data into a workflow that, in turn, feeds an analytics product or platform.

While your team has done this hundreds of times, no two clients' data is ever the same. And somehow the process of getting the client data in a specific format or schema never seems to go according to plan.

When all systems are go, and a new project is full speed ahead, why is integrating your client's data the biggest bottleneck?

Whether you call this data onboarding, data integration, data implementation, or otherwise, this is a problem universal to organizations who need to take raw data and make it fit for purpose.

This eBook explores why integrating new client data can be a stumbling block for analytics providers. It examines the shortcoming of the two most common approaches to data onboarding: Excel and coding. It explains how self-service data preparation is a far superior approach to data integration by helping analytics providers save time up front, and automate processes to deliver continuous value to their clients.

Excel and Coding: Two Approaches that Fall Short

There are two traditional approaches when it comes to integrating new client data: Excel or code. Each has its unique advantages, and both have shortcomings for data integration.

The Shortcomings of Excel: Too Manual

Who doesn't know Excel? It's ubiquitous. Most data-savvy professionals are proficient with the spreadsheet tool, and your team likely knows their way around macros, functions, and pivot tables.

But analytics providers need more out of Excel than it can give. How can you get the 30,000-foot view of your client's dataset? What's in the dataset? Does the data contain the fields and records my team needs? What are the outliers, anomalies, and missing values? Excel doesn't provide much help in answering these questions.

Once you start manipulating data in Excel, its limitations and manual tedium become obvious:

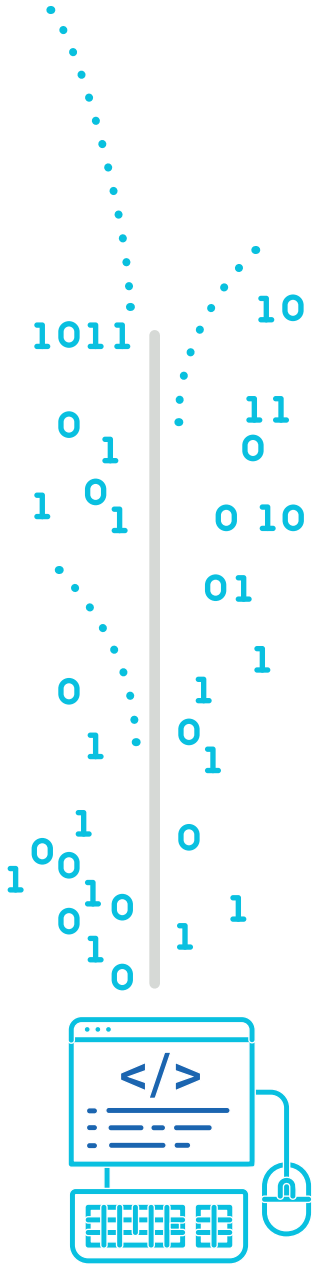
- The size of datasets exceed Excel's capacity. You can't even open large datasets, let alone use Excel functions to manipulate data without the tool timing out or crashing.
- Excel doesn't help your team avoid data quality errors resulting from manual data entry/manipulation.
- Excel doesn't generate scripts that can be automated and reused.
- Macros are too rigid for the variability in unknown client data.

Excel doesn't easily scale to meet your business needs. Since it lacks automation capabilities, each time your client services manager adds a new client to her customer list, it adds an additional client for which she must continuously prepare data. You'd like to have each member of your team managing more than a few customers each, so where do you turn? If your looking for automation, the obvious choice would seem to be code.

The Shortcomings of Code: Too Technical

While coding tools like Python or R fill in some of the gaps left open by Excel, they present their own challenges when it comes to integrating data.

Once a data integration process has been built, your team can create a script to run your Python code repeatedly to feed your data products. This automation saves your client success managers from having to go through the tedious process of preparing data each time your existing clients



need updated insights on new data. But you can't use the same script for each new client because of the variations in each client's data —naming conventions, sources of data, attributes, to name a few. Building a one-size-fits-all Python script or ETL process is a no-go.

A one-size-fits-all Python script or ETL process is no match for integrating highly variable client datasets.

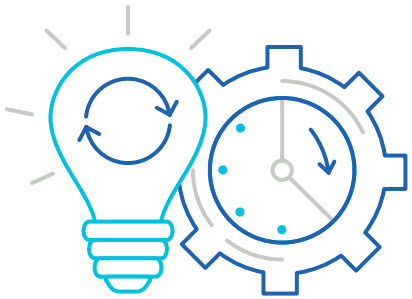
Does that rule out code entirely? Certainly not. Your implementation team or customer success team can still use Python or SQL to build a new script for each new client dataset. But how scalable is this approach, given resource limitations and the technical expertise required? Code-savvy client services managers are few and far between. Some analytics providers turn to the engineering team at this point, and ask them to build code and automation for each new client that comes on board. These companies quickly realize that this ends up taking more time than expected, and is not a good use of the engineering teams time.

Even if your company were able to recruit and hire enough code-proficient resources, those resources are human. They'd make mistakes. They're subject to the enormous time pressures of having to develop code, visualize the results through a head sample or a more robust visualization tool to validate the output, and then edit and debug the code when mistakes inevitably surface on the first go. Sure, code can be automated, but it's time consuming to build the initial scripts and difficult to debug when those initial scripts show errors. Manual code is not the droid you're looking for.

Why Self-Service Data Preparation is Just Right

When Excel is too limited, and code is too technical, what do you do? There is a better way.

It's well understood that 80% of data workers' time is spent preparing data.

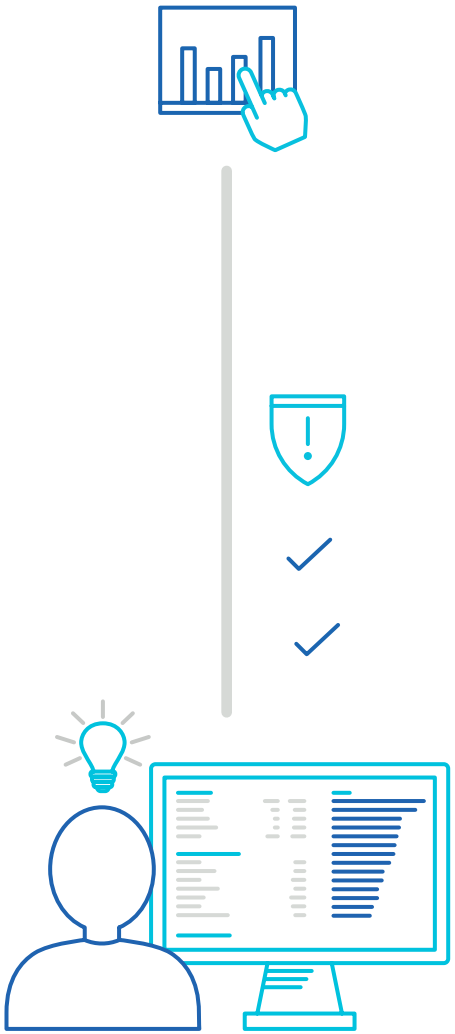


You need a self-service data preparation platform. Such a platform would empower your team to work directly with the data and build out repeatable processes quickly so you can deliver value to your clients, earn their trust, and build your business. This platform should:

- Combine the ease-of-use of tools like Excel with self-service automation capabilities
- Offer visual guidance to help users discover and understand the contents of their data—which is particularly important when working with a new client's unfamiliar data
- Speed up the time it takes to integrate a new client's data the first time—and every time thereafter by automating the work

Trifacta Saves Time in Design and Production

Trifacta helps analytics providers like yours more efficiently onboard client data. Trifacta's self-service data preparation platform helps organizations explore, transform and join together diverse data. Whether you're working with files on your desktop, disparate data in the cloud or within large-scale data lake environments, Trifacta accelerates the process of getting data ready to use.



See Your Work

Trifacta saves analytics providers time in both the design stage of data preparation pipelines and in putting them into production. When you load an unfamiliar dataset from a data warehouse or data lake, file system, or other data source into the Trifacta platform for the first time, you get immediate visual feedback about your data. Column histograms and data quality bars point you to information about the contents in each column and indicate whether the values are consistent and of high quality, or if they have errors to address.

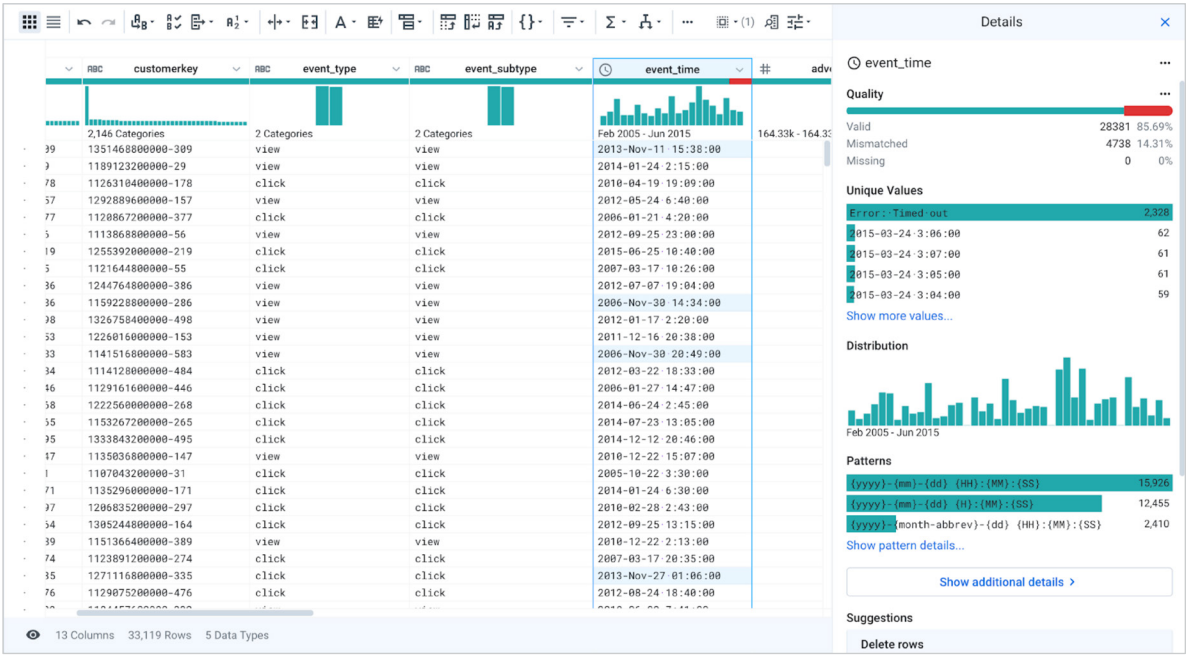
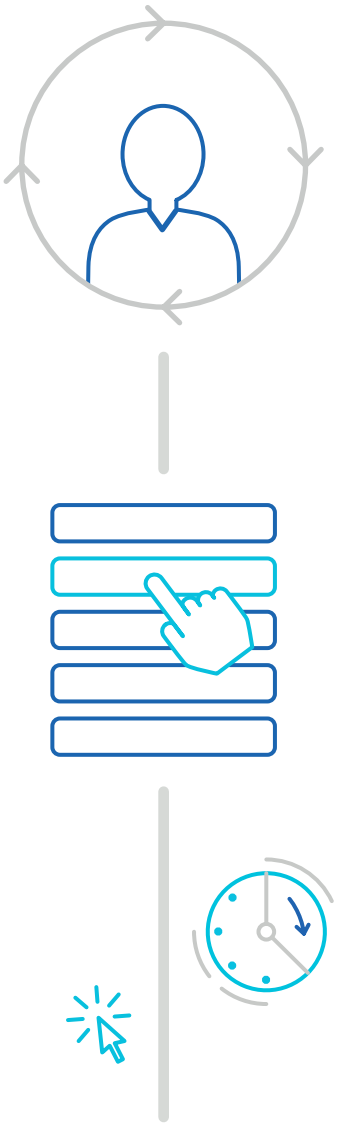


Fig 1. With the Trifacta self-service data preparation platform, you get immediate visual feedback about the quality of your data.



If you have an existing data format you need the data to conform to, the platform's Rapid Target feature adds an additional layer of visual guidance, overlaying that output format you need to achieve and helping to guide your transformations to get there.

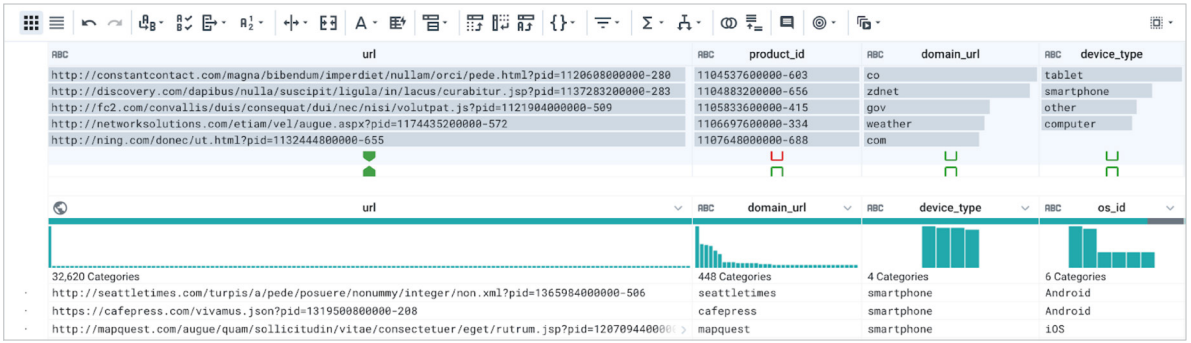


Fig 2. The Trifacta platform's Rapid Target feature overlays your desired output formats and guides you through the transformations needed to achieve them.

Providing real-time previews of every step gives your team confidence in the steps they are applying, getting it right the first time, and saving valuable time that would otherwise be spent debugging and searching for the cause of errors.

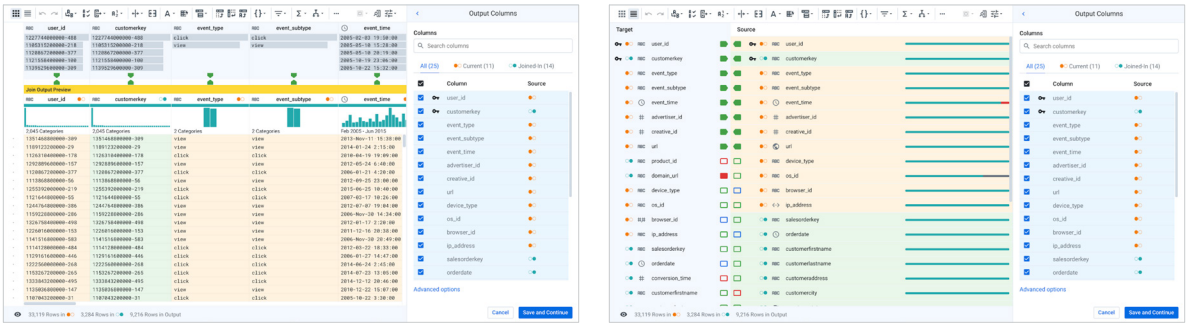
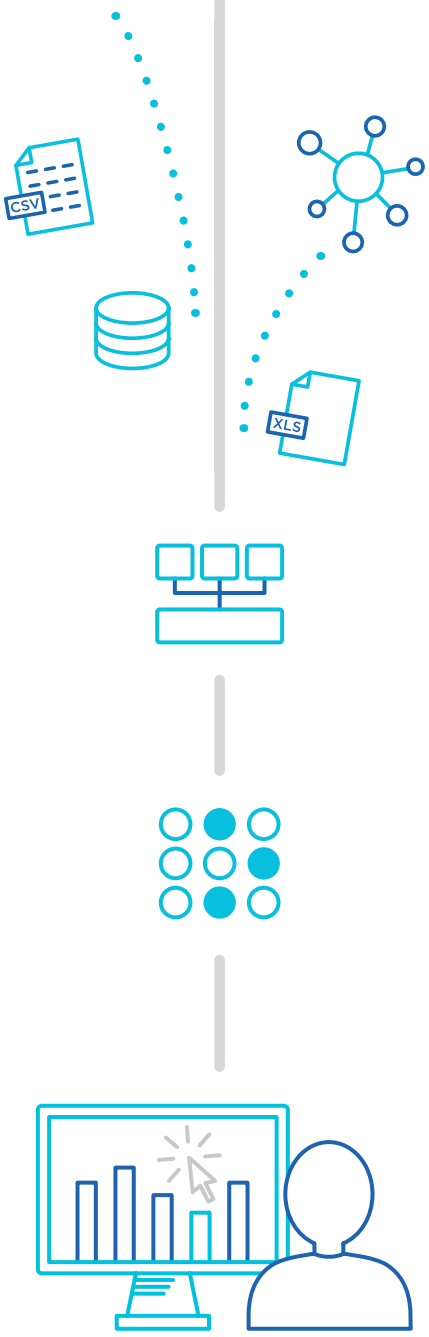


Fig 3. With the Trifacta platform, blending data is easy. Your team gets suggested columns to join on, and visual previews of the join. You can preview the join in a grid view (left) or a columns view (right).



Clean and Structure Your Data

Based on column data types, the Trifacta platform provides intelligent suggestions for cleaning up formatting issues. Cleaning involves taking out data that might disrupt the onboarding process. A null value, for example, might bring integration to a screeching halt; it may need to be replaced with a zero or an empty string. Particular fields may need to be standardized by replacing the many different ways that a state for example might be written out —such as CA, Cal and Calif, or 2013 Nov 11 and 11-11-2013—with a single standard format.

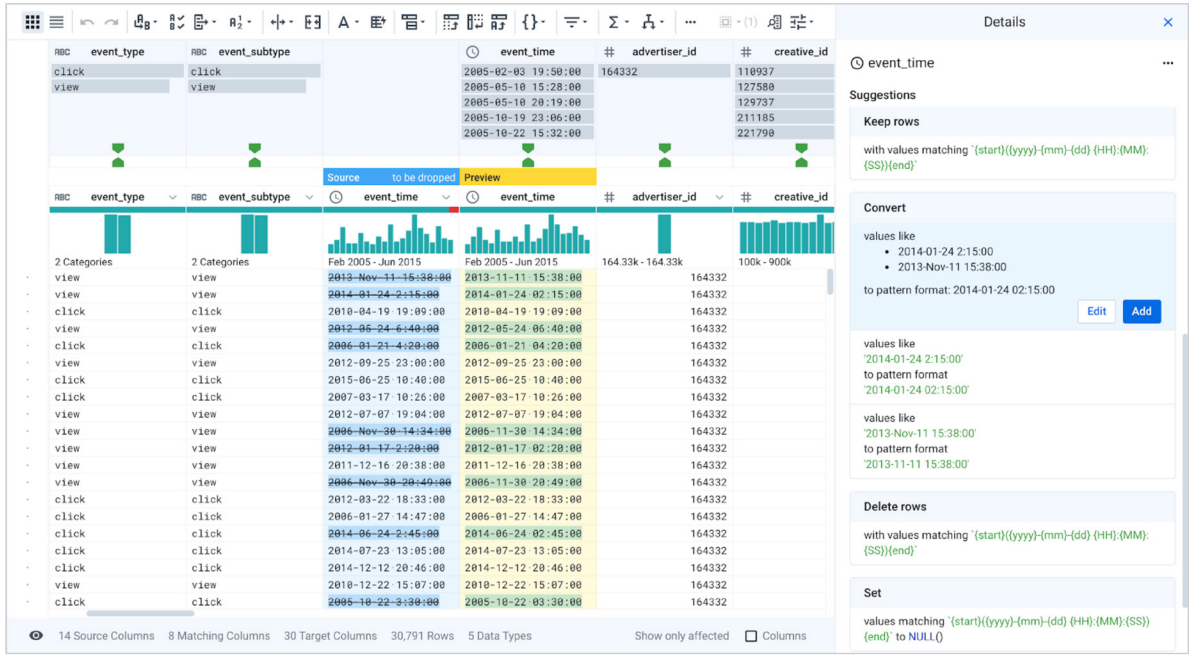


Fig 4. The Trifacta Platform offers suggestions for cleaning up formatting issues, and gives you real time previews of what the step accomplishes.

Since raw client data comes in many shapes and sizes, you need to give it structure. This can mean creating columns and rows from particularly unstructured datasets, extracting important information, flattening arrays into individual rows or unnesting objects into separate columns. The Trifacta platform makes it easy to work with all types of data, from free form text files to highly structured CSVs.

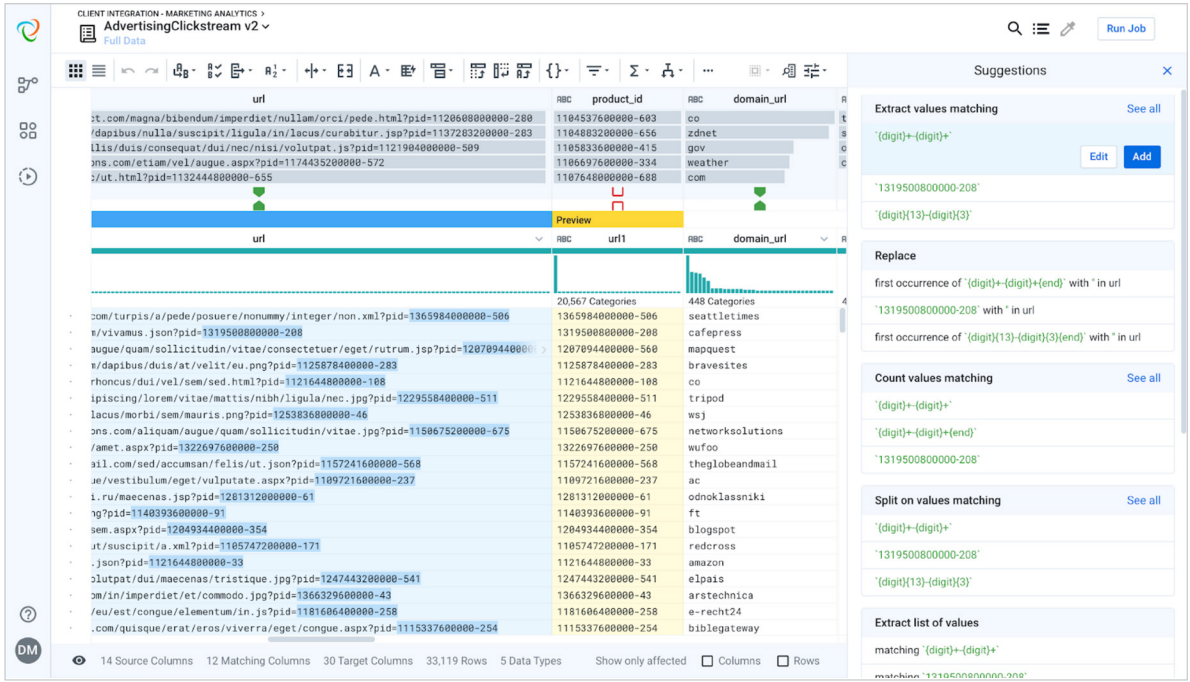
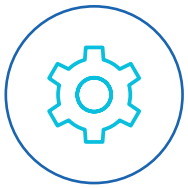
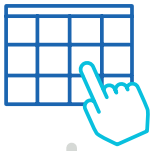
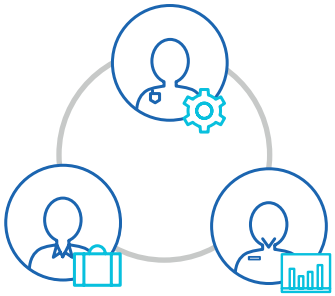


Fig 5. The Trifacta Platform makes it easy to extract information as part of integrating client data.







Automate the Onboarding Process

Trifacta can help accelerate the design time for technical and nontechnical users alike, but the real value is in automation. Once you have created the steps needed to take your clients data from raw to refined, the Trifacta platform allows you to schedule and orchestrate job runs, so you can make sure new output is generated each time you need to provide value to your clients. Your teams can also receive alerts through email and third-party applications like Slack whenever a job completes, providing full transparency into each user's workflow. As an admin, you can view all of the scheduled jobs of all of your users to monitor your whole client portfolio and ensure they are each getting updated when appropriate.

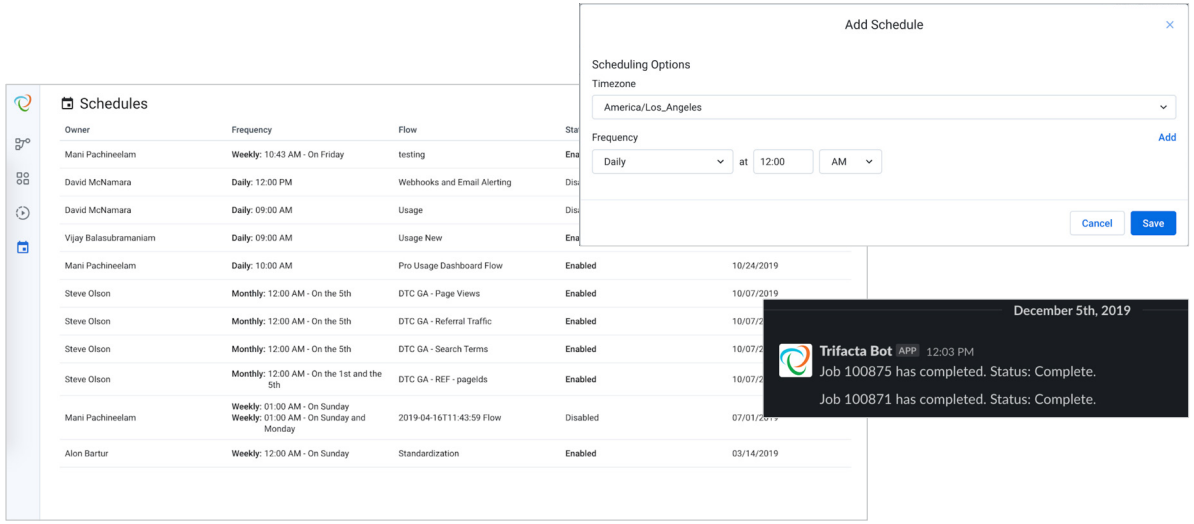
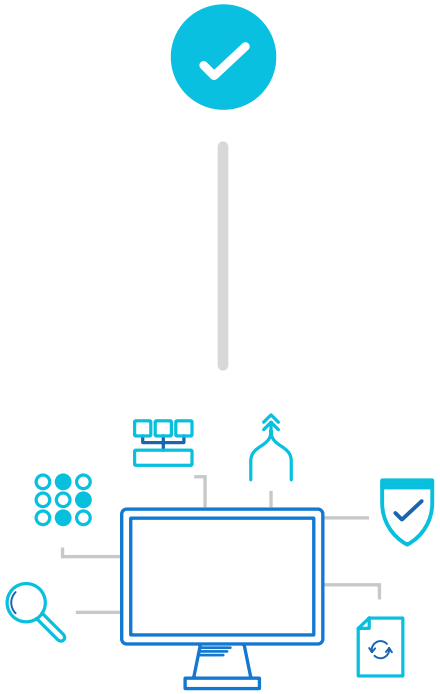


Fig 6. With the Trifacta platform, job runs can be scheduled and orchestrated to ensure new outputs are generated each time you need to provide value to your clients. You can receive alerts when jobs are completed through email or third party applications like slack. As an admin, you can view all of the scheduled jobs of all of your users to monitor your whole client portfolio and ensure they are each getting updated when appropriate.



Conclusion

The Trifacta self-service data preparation platform combines ease-of-use for faster client data integration with automation for continuing value, saving your team time and unleashing the productivity of your client service managers.

The platform is fast and easy to set up and configure. As a SaaS platform, it requires no installation or management from your end.

Interested in seeing the benefits for yourself? Try Trifacta for free today!

Follow Trifacta

@Trifacta Trifacta Trifacta

Ready to Get Started?

If you're ready to start that journey, sign up for a free trial of the leading modern data wrangling solution from Trifacta and start wrangling your data with just a few clicks at www.trifacta.com/start-wrangling.

Schedule a demo and learn more about our data preparation solution.

TRIFACTA®

575 Market St, 11th Floor
San Francisco, CA 94105
1 844 332 2821
www.trifacta.com