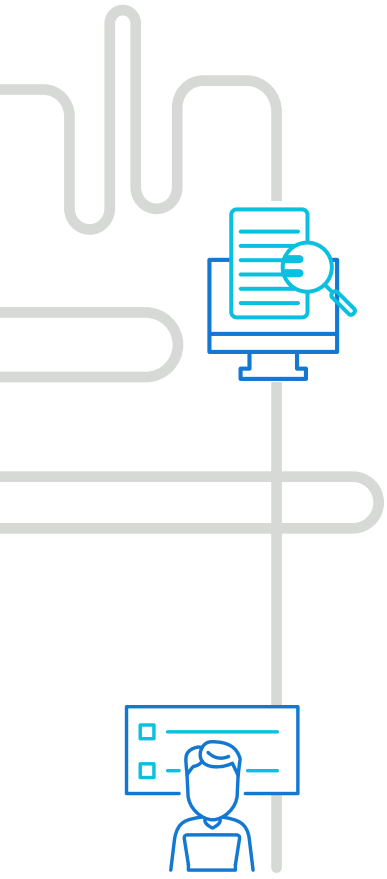
A New Approach to **Clean Data**

Understanding the five tenets of proper data preparation

TRIFACTA®

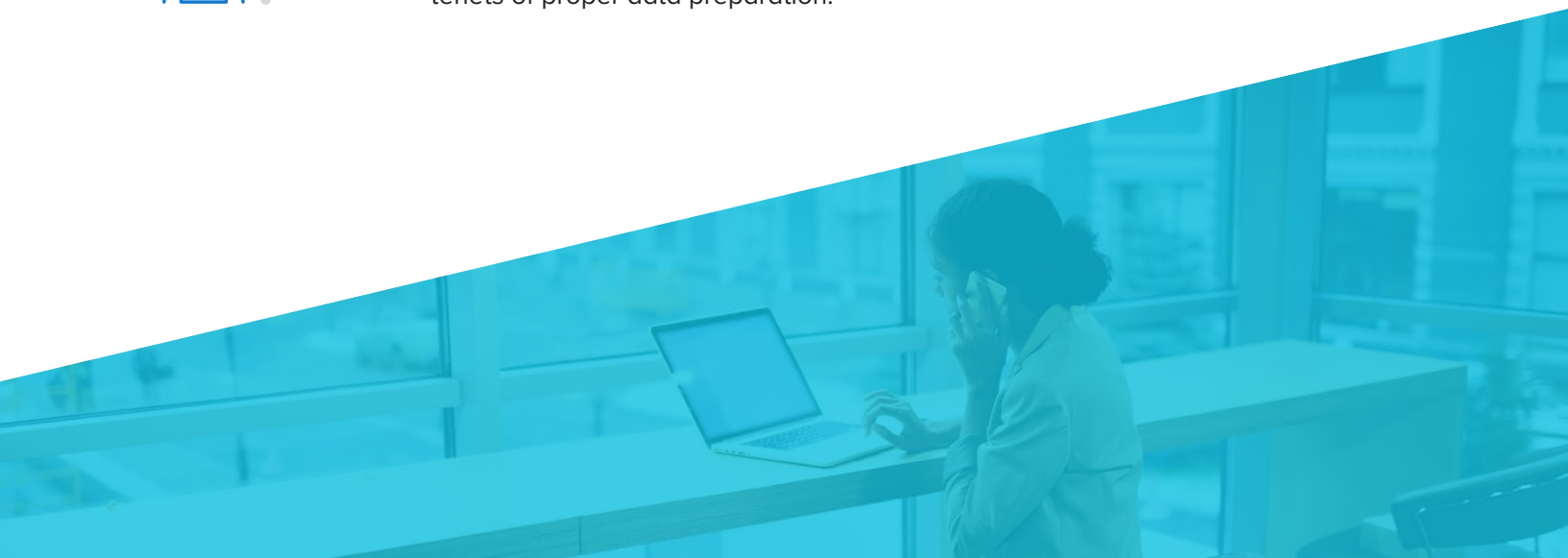
The data industry is undergoing a massive evolution.



Over the past decade, the analytics technology stack has changed dramatically, with new processes and technologies informing how data is stored and processed. One of the key motives for this change has been a heightened focus around improving the accessibility of data. This has called for a shift away from rigid, IT-led processes toward a model that prioritizes self-service in collaboration with IT. Democratizing data, instead of limiting it to a small group of technical resources, is an increasingly important strategy to spawn new and advanced analytics initiatives.

Amidst this change, the bottleneck of data preparation has risen as the most critical element of the analytics and data management process in need of new technology and process. It's been widely covered that getting data cleaned and prepared for analysis takes up to 80 percent of the time and resources in any data project—making it the biggest area of inefficiency when working with data, but also the biggest area for improvement. There's no avoiding how critical data preparation is to deriving value from data. Clean data is essential to being able to stand behind the validity of your analysis and this is a key reason why new technologies and approaches to data preparation have been shown to have the biggest impact on improving analytics efficiency.

Understanding the changes taking hold to data preparation technology are essential, as is the new best practices that come with effectively utilizing these technologies. Below, we emphasize the criticality of data preparation, what to look for in modern solutions, and why these changes mandate new thinking around organizational processes for preparing data—or what we call the five tenets of proper data preparation.



The data on data preparation isn't pretty

In taking a closer look at the traditional approaches to data preparation, it's clear that they are a major drain on resources. We found in a survey earlier this year that as many as 60 percent of IT professionals spend half or more of their time at work on data quality assurance, cleanup or preparation. Based on Glassdoor salary estimates and IDC's estimation that there are 18 million IT operations and management professionals globally, that adds up to a price tag of over **$450 billion** spent on data preparation by organizations around the world.

In addition to the hard costs, data janitorial work sinks morale as well. Unsurprisingly, both IT professionals and data analysts would rather be doing something else. Sixty percent of IT professionals believe they are overqualified to spend as much time as they do on data preparation. Many believe they could bring more value to their organizations if their time were spent modelling, finding insights or designing programs.

Data analysts are also spending too much time preparing data:

92 percent would choose to focus on another analytic activity rather than data preparation, yet

65 percent are spending at least half their time preparing data for analytic use.

Employing manual methods could be hurting not only your efficiency and staff satisfaction, but an elevated risk of error as well, or of losing potentially critical insights or conclusions as a result of mistakes made while preparing data.

The advent of data preparation platforms

No longer must IT professionals fill the role of "head janitor" in data preparation. The work of integrating and cleaning disparate data from various sources to prepare it for analysis is best done by those who understand the data and how it needs to be used. It's tedious work—or it used to be—but that's changing with the advent of visual, intelligent data prep solutions. Solutions that can accelerate the process and reduce errors.

More and more organizations are adopting new machine learning-driven solutions to support their data preparation needs, and transitioning away from manual approaches using Excel or code. And so, we find ourselves in the midst of a major shift. Organizations increasingly recognize the need for more automated approaches that put the end user—the analyst—in the driver's seat when it comes to preparing the data, preparing it the right way for the intended use case, and ultimately getting the data ready to use faster.

We're moving to a new normal in data analysis that means adopting a whole new approach to the standardization and cleansing of data for analytic use.

Understanding the five tenets of proper data preparation

This is why we at Trifacta developed what we call the Clean Data Manifesto, as a call to action to anyone who works with data. The Clean Data Manifesto is about committing to clean data, and committing to best practices around how to prepare and work with data that can optimize your results for analysis. As part of this manifesto, we have identified five tenets of proper data preparation practice that we consider to be critical:

Clean Data Tenet #1: Prioritizing and Setting Targets

Context matters: The definition of "good" data varies from project to project, even for the same data. An in-depth understanding of your use case will ultimately determine the data quality issues that matter most and what "good enough" looks like.

What data is most essential to the success of your project? What level of quality is really necessary? How significant are the risks of bad data? If you don't know up front what is important to your use case, then you risk wasting resources and minimizing your return on effort (RoE).

That's why it's so important that the people who know the data best and understand the context do this work themselves. Putting that power in the hands of analysts will help your organization more efficiently prepare the data—and ultimately produce analyses founded on data that is clean, suitable and appropriate.

Clean Data Tenet #2: Identify Issues Early and Often

Confirm your data is sound by keeping the 4 Cs of data quality as you prepare your data: the consistency, conformity, completeness and currency of the data.

The truth is, the best way to get ahead of issues is to identify them early and often. This is difficult to do when you're shipping your data requirements off to IT, and then waiting to receive it back before being able to assess and define new requirements. Or when you're trying to scroll through rows upon rows of Excel sheets or code—these time-consuming tasks and cycles don't lend well to efficient detection.

Using intelligent tools to assess the data through both statistical and empirical approaches will help you uncover any anomalies or Cs that don't pass muster, and focus on refinement efforts.

Clean Data Tenet #3:

Collaborative Curation

Data preparation is a team effort that requires the seamless orchestration of a lot of moving parts in order to curate data quality. Leveraging external, third-party data sets is often critical to enhancing your analysis.

Collaboration strengthens your data preparation by bringing a broader collective context to the effort. In order to build a collaborative environment, it's important to look for alternatives to tools that limit transparency, such as scripting languages or common spreadsheet tools.

Modern data preparation platforms like Trifacta operate off of a visual-based interface, which allows anyone in the organization to speak the same language, and maintains clear data lineage to improve transparency and encourage feedback about how particular data sets have been transformed. These platforms support and promote the sharing of data, data preparation recipes and entire workflows so users can feed off of the work of their peers and stronger collective intelligence across their organization.

Clean Data Tenet #4:

Constantly Monitor

You can't leave everything to automation. Modern data preparation—like so many modern business practices—is enabled by the marriage of automated and manual processes. You should always stay vigilant about the quality or "cleanliness" of your data.

Combining technology with human monitoring is the only way to arm analysts with better process efficiency and true quality assurance—so they can produce results that are trustworthy without creating new issues in the process.

Clean Data Tenet #5:

Ensure Transparency

The ability to trust your data hinges on trusting the process you use to clean it. This means having a full audit trail to understand lineage and chain of custody. This audit trail also must be in a format that a range of individuals with varying skill sets can understand. Why? Because it's not enough to just communicate your results—you need to show your work.

This is critical for both external compliance requirements and for your own internal credibility. To ensure your results can be reproduced, understood and trusted, you have to be able to audit how and when the data was transformed, as well as who transformed it.

To build trust, ensure consistency and remove potential bias, be transparent about the ways in which you've altered the data.

Adopting a new approach to data preparation

Following these tenets is critical to producing results your organization can trust and stand behind, and to doing so efficiently. And getting to that point means using modern data preparation solutions to power your analytics processes from start to finish.

What can you expect from transitioning to this new normal?

First: IT no longer needs to be the bottleneck by handling all of the hands-on prep work, and can instead focus on governance, security, and scaling processes across the organization. Second: and perhaps most importantly: you empower analysts to take the reins (in a governed, secure manner), to prepare data in a way that's both efficient and appropriate for the use case at hand, and ultimately, to stand behind the integrity of their analysis.

#CommitToCleanData #NoExcuses