+ a b l e a u

Dirty data is costing you

4 ways to overcome common data prep barriers

Introduction

If you've ever analyzed data, you know the pain of digging into your data only to find that the data is poorly structured, full of inaccuracies, or just plain incomplete. You're stuck adapting the data in Excel or writing complex calculations before you can answer a simple question.

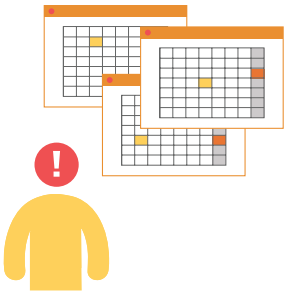
Data preparation is the process of getting data ready for analysis, including data discovery, transformation, and cleaning tasks—and it's a crucial part of the analytics workflow. A recent Harvard Business Review study reports that people spend 80% of their time prepping data, and only 20% of their time analyzing it. And this statistic isn't restricted to the role of the data stewards. Data prep tasks have bled into the work of analysts and even non-technical business users.

Even those who aren't directly performing data preparation tasks feel the impact of dirty data. The amount of time and energy it takes to go from disjointed data to actionable insights leads to inefficient ad-hoc analyses and declining trust in organizational data. These slower processes can ultimately lead to missed opportunities and lost revenue. In fact, Gartner research shows that the "average financial impact of poor data quality on organizations is $9.7 million per year."*

Table of contents

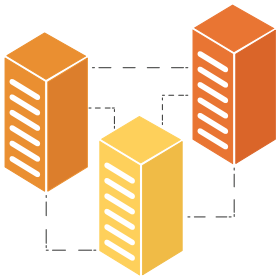
Why dirty data happens

Enterprises are taking steps to overcome dirty data by establishing data catalogs and glossaries. But even with these practices, it is likely for some level of dirty data to seep through the cracks of day-to-day operations. Dirty data commonly happens due to:



1. Human error

It is the most common cause of dirty data, according to Experian. Errors can pop up in a variety of ways, from variability in data entry practices to employees manually inputting values into spreadsheets. Even a simple spelling error could pose challenges down the line when someone needs to analyze the data.



2. Disparate systems

Organizations often store data in several disparate systems that have different structures and requirements. When it comes time to integrate this data, analysts are left with duplicate or missing fields or inconsistent labels. Data fields or values might also have the same meaning, but use different names or values across systems. These issues get even trickier when companies need to bring in data from external sources like partners or vendors, where it could be encoded differently or aggregated at different levels.

3. Changing data requirements

Businesses evolve and as a result, data administrators and engineers need to make changes to the data—changing its granularity, deprecating fields if they're not being used, or introducing new fields as needed. Although these changes are necessary, they are not always widely communicated across a business, and analysts may not even know about these changes until they bring the data into a self-service BI or data prep tool.

Four common data prep issues (and how to solve them)

01 **Issue 1:** Rigid and time-consuming processes don't keep up with demand

Analysts report that the majority of their job is not analysis, but cleaning and reshaping data. This can occur with an ETL process, in a self-service data prep tool like Alteryx or Trifacta, or in spreadsheet tools like Microsoft Excel. Every time new data is received, analysts need to repeat manual data preparation tasks to adjust the structure and clean the data for analysis. This ultimately leads to wasted resources and an increased risk for human error.

Beyond the frustration of messy data, both analysts and business users struggle to even access the data they need. Traditionally, data preparation has lived within IT—and only certain teams have the ability to bring new data sources into a centralized data warehouse. Those who don't have this ability either conduct their own data prep in programs like Excel or wait for another team to do it for them. Cathy Bridges, Tableau Developer at SCAN Health Plan noted that "When we need to make changes to a data set, it can take weeks at a minimum and often months."

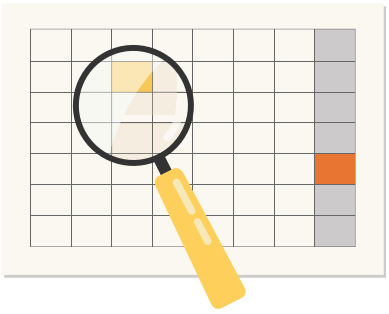
> When we need to make changes to a data set, it can take weeks at a minimum and often months.
>
> — Cathy Bridges, Tableau Developer, Scan Health Care

Solution: Develop agile processes with the right tools to support them

Many organizations are adopting self-service data preparation solutions for exploration and prototyping. Self-service data preparation tools put the power in hands of the people who know data the best—democratizing the data prep process and reducing the burden on IT. "The added value of a self-service data prep tool is that everyone can become a master of the data," said Venkatesh Shivanna, Senior Data Analytics Manager and Architect at a popular gaming company. "Analysts can do the ad-hoc data cleansing tasks themselves instead of waiting in a queue."

Every organization has specific needs and there is no 'one-size-fits-all' approach to data preparation, but when selecting a self-service data preparation tool, organizations should consider how the tool will evolve processes towards an iterative, agile approach instead of creating new barriers to entry. Jason Harmer, consultant in IT process management at Nationwide Insurance explained how "you can't really democratize data without letting people understand the full data prep process. Visual data prep allows people to see the full end-to-end process, seeing potential flags earlier on—like misspellings in the data, extra spaces, or incorrect join clauses. It also increases confidence in the final analysis." People will have a greater desire to prepare and understand their data if they can see how the impact of their data prep steps.

02 **Issue 2:** Data preparation requires deep knowledge of organizational data



Before preparing data, it is crucial to understand its location, structure, and composition, along with granular details like field definitions. Some people refer to this process as "data discovery" and it is a fundamental element of data preparation. You wouldn't start a long journey without a basic understanding of where you're going, and the same logic applies to data prep.

The emergence of self-service BI and its drag-and-drop functionality has made data discovery easier for business users, providing them with a deeper knowledge of the existing structure and contents of their data sets. But because of information silos, these users often have less insight into the entire data landscape of their organization—what data exists, where it lives, and how it is defined. Confusion around data definitions, for example, can hinder analysis or worse, lead to inaccurate analyses across the company. For example, if someone wants to analyze customer data, they may find that a marketing team might have a different definition for the term "customer" than someone in finance.

Solution: Create company standards for data definitions

Visual, self-service data prep tools allow analysts to dig deeper into the data to understand its structure and see relationships between tables. Because they can understand the profile of their data, analysts can easily spot unexpected values that need cleaning. Although this technology brings clarity to the data, people will still need support from others in their company to understand details like field definitions.

One way to standardize data definitions across a company is to create a data dictionary. A data dictionary helps analysts understand how terms are used within each business application, showing the fields are relevant for analysis versus the ones that are strictly system-based. Brian Davis, Project Engineer at an energy company calls data dictionaries "invaluable." Brian says, "I regularly combine data from accounting with data from field technicians. Defining the initial data along with calculated fields drives more accurate analyses and reduces the amount of time spent determining which field or table to use."

Developing a data dictionary is no small task. Data stewards and subject matter experts need to commit to ongoing iteration, checking in as requirements change. If a dictionary is out of date, it can actually do harm to your organization's data strategy. Communication and ownership should be built into the process from the beginning to determine where the glossary should live and how often it should be updated and refined.

03 **Issue Three:** "Clean data" is a matter of perspective

Different teams have different requirements and preferences regarding what makes for "well-structured" data. For example, database administrators and data engineers prioritize how data is stored and accessed—and columns may be added that are strictly for databases to leverage, not humans. When an engineer builds a data warehouse specifically for analysis, they prioritize the core business metrics that answer the majority of questions. If the information that data analysts need isn't already in the data set, they may need to adjust aggregations or bring in outside sources. This can lead to silos or inaccuracies in the data.

Cathy Bridges, Tableau Developer at SCAN Health Plan, explained how analysts often have to go back and update a data set that has already been cleaned by another team. "Bringing in additional columns can be a long and arduous process. For example, if I need totals versus breakout, I need to duplicate the data source—and it can be a pain."

> "A data prep tool should equip the one-off questions from the analysts and also be repeatable. When I build out the logic, it's saved in a file somewhere. And the next time, I can reopen that same file, repoint at the same data sources and start from where I left off in that workflow."
>
> — Gordon Strodel, Information Management And Analytics, Slalom

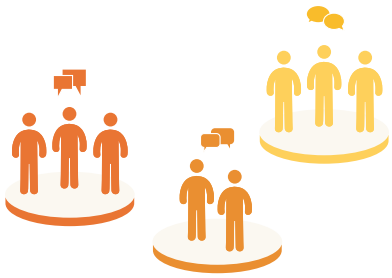
Solution: Put the power in the hands of the data experts

Self-service data prep gives analysts the power to polish data sets in a way that matches their analysis, leading to faster, ad-

hoc analyses and allowing them to answer questions as they appear. It also reduces the burden on IT to restructure the data whenever an unanticipated question arises. This can also reduce the amount of duplicated efforts because other analysts can reuse these models. If the datasets are valuable on a wide scale, you can combine them into a canonical set in the future.

"A data prep tool should equip the one-off questions from the analysts and also be repeatable," says Gordon Strodel, Information Management and Analytics Consultant at Slalom. "When I build out the logic, it's saved in a file somewhere. And the next time, I can reopen that same file, repoint at the same data sources and start from where I left off in that workflow."

04 Issue Four: The hidden reality of data prep silos



Advanced data preparation tools can be complex, which means this capability is often restricted to a select number of power users. But even if analysts and business users don't have access to data preparation tools, it doesn't mean that they aren't already performing these tasks in other applications. Self-service business intelligence tools have opened up data analysis capabilities to every level of user, but in order to get insights into their data, these users still need to rely on IT for well-structured data. Instead of waiting days or months for the data, users extract data from systems and prepare their data in spreadsheets. The result is a newly structured data set that serves a singular purpose and departments often duplicate efforts without even knowing it. This process leads to an abundance of data silos, which aren't efficient, scalable, or governed.

> "Data dictionaries are invaluable. I regularly combine data from accounting with data from field technicians. Defining the initial data along with calculated fields drives more accurate analyses and reduces the amount of time spent determining which field or table to use."
>
> — Jason Harmer

Solution: Create consistency and collaboration within the data prep process

Combatting silos starts with collaboration. Survey research from the Business Application Research Center (BARC) showed that the companies that were most satisfied with their data prep processes were the ones that "made data preparation a

shared task between IT and business departments." Jonathan Drummey, Consultant at DataBlick and Data Visualization Specialist at PATH explained that throughout this process, there should be "people downstream and someone (or multiple people) upstream. The upstream people are taking feedback from the downstream people to do cleanup, usually around data quality issues and availability of supplemental data sets."

Adopting a self-service data prep across an organization requires users to learn the ins and outs of the data. Since this knowledge was historically reserved for IT and data engineering roles, it is crucial that analysts take time to learn about nuances within the data, including the granularity and any transformations that have been done to the data set. Scheduling regular check-ins or a standardized workflow for questions allows engineers to share the most up-to-date way to query and work with valid data, while empowering analysts to prepare data faster and with greater confidence.

About Tableau

Tableau is the enterprise analytics platform that helps people see and understand data. Give people access to intuitive visual analytics, interactive dashboards, and limitless ad-hoc analyses that reveal hidden opportunities and eureka moments alike. Get the security, governance, and management you require to confidently integrate Tableau into your business application and deliver the power of embedded analytics at scale.

tableau