

alteryx

E-BOOK

5 Steps to a Bulletproof Data Prep Strategy



- 03 Setting up for Success
- 03 Data Prep 101
- 04 Data Blending
- 05 Data Exploration
- 06 Data Cleansing
- 07 Data Profiling
- 08 Extract, Transform, Load (ETL)
- 09 Getting Faster, Smarter Insights
- 09 Automating Data Prep with Alteryx
- 11 Why Data Professionals Love Alteryx



Setting up for Success: The Importance of Data Prep

Data prep may not be glamorous, but it's the structural foundation of good business analysis. If you don't clean, validate, and consolidate your raw data the right way, you won't be able to get meaningful answers.

But in a typical organization, data lives in silos, where its potential can't be fulfilled — and in spreadsheets, where it's manipulated by hand. The more data you acquire and the more complex it gets, the more these problems amplify.

What if you could automate some of these processes and make them fast, transparent, and repeatable?

Understanding the Fundamentals

Before we talk solutions, let's take a closer look at what you should be planning when it comes to data preparation. A successful approach to data prep includes these functions:

DATA BLENDING

Combining data from multiple sources into a single data set for analysis and visualization.

DATA EXPLORATION

Evaluating the data's quality, examining its distribution, and analyzing the relationship between each variable to better understand how to compose an analysis.

DATA CLEANSING

Improving quality and overall productivity by deleting unnecessary data, removing poor quality data, or fixing inaccuracies to craft error-proof insights.

DATA PROFILING

Spot poor-quality data before it negatively impacts your results.

TRANSFORMING DATA

Formatting, orienting, aggregating, and enriching the datasets used in analysis to produce more meaningful insights.

DATA PREP STEP 1: Data Blending

The more high-quality sources you incorporate into your analysis, the deeper and richer your insights. Typically, any project you undertake will require six or more data sources — both internal and external — requiring data blending tools to fuse them together seamlessly.

The moment before blending is kind of like looking over the edge of a cliff. What if you introduce a new dataset and it sets off an avalanche of compatibility issues, and you can't undo the damage? Sometimes the complexity of the work makes it tough to be completely confident in the results. It's always better to have a solution that allows you to go back in time to the point before you made changes.

DATA BLENDING USUALLY INVOLVES THREE STEPS:

ACQUIRE AND PREP

If you're using modern data tools rather than trying to make files conform to a spreadsheet, you can include almost any file type or structure that relates to the business problem you're trying to solve and transform all datasets quickly into a common structure.

BLEND

In spreadsheets, this is where you flex your VLOOKUP muscles. If you're using self-service analytics instead, this process is as simple as drag-and-drop.

VALIDATE

It's important to review your results for consistency and explore any unmatched records to see if more cleansing or other prep tasks are in order.

DATA PREP STEP 2: Data Exploration

Before you start working with a new dataset, it's a good idea to step boldly into the raw material and do a bit of exploring. Although you might start with a mental picture of what you're looking for, or a question you'd like to see answered, it's best to keep an open mind and let the data do the talking.

Data exploration used to require the code-writing skills of IT engineers, which amounted to a locked gate between raw data and the people who analyzed it. But now, by using automated tools as building blocks throughout the data prep process, data analysts and business users can plunge right into a dataset themselves and explore whatever lies within.

DATA EXPLORATION TECHNIQUES THAT DELIVER IMPACTFUL RESULTS:

- ✔ Scan column names and field descriptions to see if any anomalies stand out, or if any information is missing or incomplete.
- ✔ Do a temperature check to see if your variables are healthy. How many unique values do they contain? What are the ranges and modes?
- ✔ Spot any unusual data points that may skew your results. You can use visual methods — box plots, histograms, or scatter plots — or numerical approaches such as z-scores.
- ✔ Scrutinize those outliers. Should you investigate, adjust, omit, or ignore them?
- ✔ Examine patterns and relationships for statistical significance.

DATA PREP STEP 3: Data Cleansing

Your analysis is only as good as the data that powers it. That's why data full of errors and inconsistencies carries a hefty price tag — studies show that dirty data can shave millions off a company's annual revenue.

To prevent significant losses, it's critical to scrub your dataset until it sparkles. Data cleansing can be done manually, but it's a tedious and time-consuming process. Automated data cleansing tools, on the other hand, can do most of this work with a few clicks.

THE FOLLOWING STEPS ARE COMPLETED IN THE CLEANSING STAGE:

- ✔ Ditch duplicate records that clog your server space and distort your analysis.
- ✔ Remove any rows or columns that aren't relevant to the problem you're trying to solve.
- ✔ Investigate and possibly remove missing or incomplete information.
- ✔ Fix structural errors — typography, capitalization, abbreviation, formatting, extra characters.
- ✔ Validate that your work is accurate, complete, and consistent, documenting all tools and techniques you used.

DATA PREP STEP 4: Data Profiling

Data profiling is similar to data exploration but with a more intense focus. Profiling data means examining a dataset specifically for its relevance to a particular project or application. Profiling determines whether a dataset should be used at all — a big decision with potential consequences for the business.

Data profiling is also intricate and time-consuming, which is why many companies use automated analytics tools to avoid business users having to rely on data specialists.

THERE ARE THREE DATA PROFILING TECHNIQUES PERFORMED IN THIS ORDER:

STRUCTURE PROFILING

How big is the dataset and what types of data does it contain? Is the formatting consistent, correct, and compatible with its final destination?

CONTENT PROFILING

This is the stage where you'll run summary statistics on numerical fields, check for nulls, blanks, unique values, and systemic errors in spelling, abbreviations, or IDs.

RELATIONSHIP PROFILING

Are there spots where data overlaps or is misaligned? What are the connections between units of data? Identify and describe all relationships and preserve them if you move the data to a new destination.

DATA PREP STEP 5: Extract, Transform, Load (ETL)

With the enormous volume and complexity of data sources available to you, it's inevitable that you'll need to extract it, integrate it, and store it in a centralized location that allows you to retrieve it for analysis whenever you need it.

That process is known as ETL, or Extract/Transform/Load, and is the centerpiece of a modern data strategy. ETL can also help you migrate data during a disruption — such as an upgrade to a new system, or a merger with another business.

PERFORM THE FOLLOWING ETL STEPS:

EXTRACT

Pull any and all data — structured or unstructured, one source or many — and validate its quality.

TRANSFORM

Do a deep cleanse here, and make sure your formatting matches the technical requirements for your target destination.

LOAD

Write the transformed data to its storage location — usually, a data warehouse. Then sample and check for data quality errors.

Faster, Smarter Insights: The case for automation

Imagine a business environment without silos and spreadsheets. If you were able to access data in any format and automate your current prep processes with a powerful software platform, what would that look like — for you and for your organization?

Businesses that automate data preparation realize the following benefits:

QUICK WINS

Switching to an automated platform almost always produces a measurable return in a matter of days or weeks.

TIME FOR INSIGHTS

Automation completely changes the focus of an analyst's workday from menial tasks to creative ones. And you'll never have to solve the same data problem twice.

ANALYTICS UPSKILLING

When you eliminate the need for data gatekeepers, you can engage the entire organization. Employees at all levels begin coming up with new ways to expand their own capabilities.

Automating data prep with Alteryx

Alteryx is a quick-to-implement, end-to-end data analytics platform that allows you — and everyone you work with — to solve business problems faster and deliver more value.

Use our platform to discover, prep, and analyze all your data, plus deploy and share analytics at scale for deeper insights.

What's in it for you:

- Data preparation at light speed
- Repeatable workflows
- Code-free modeling through an intuitive interface
- Support for nearly every data source and visualization tool
- Performance, security, collaboration, and governance
- ROI and then some

With Alteryx, getting started with automation is easy. You can ingest and prepare data from anywhere and upskill with self-service analytics.

Why Data Professionals Love Alteryx

“It’s a Swiss Army Knife, not just an ETL tool. There are all kinds of connectors, so you can connect to a wide variety of sources.”

[READ CUSTOMER STORY](#)

Meghan Gohil, Automation and Data Visual Manager, Fender Guitar

“Doing the first 80% in analytics and the last 20% in departments that are closest to the data allows us to support the needs of the entire organization.”

[READ CUSTOMER STORY](#)

Armen Rostamian, VP of Marketing Intelligence and Analytics, BODi Online Wellness Platform

“We don’t have to spend 18 hours just on data manipulation and have more time for value-added activity.”

[READ CUSTOMER STORY](#)

Maddie Hasiewicz, CPA and Tax Technology Manager at Baker Tilly

“A powerful software like Alteryx can enable people within this space to be more nimble.”

[READ CUSTOMER STORY](#)

Alula Zeryihun, Financial Services Tax Manager, RSM

Experience Alteryx for yourself with a [30 day free trial](#).

Plus you can create value on day one with prebuilt analytic workflows using one of our [Starter Kits](#).



alteryx

About Alteryx

Alteryx powers actionable insights with the AI Platform for Enterprise Analytics. With Alteryx, organizations can drive smarter, faster decisions with a secure platform deployable in on-prem, hybrid, and cloud environments. More than 8,000 customers globally rely on Alteryx to automate analytics to improve revenue performance, manage costs and mitigate risks across their organizations. To learn more, visit www.alteryx.com