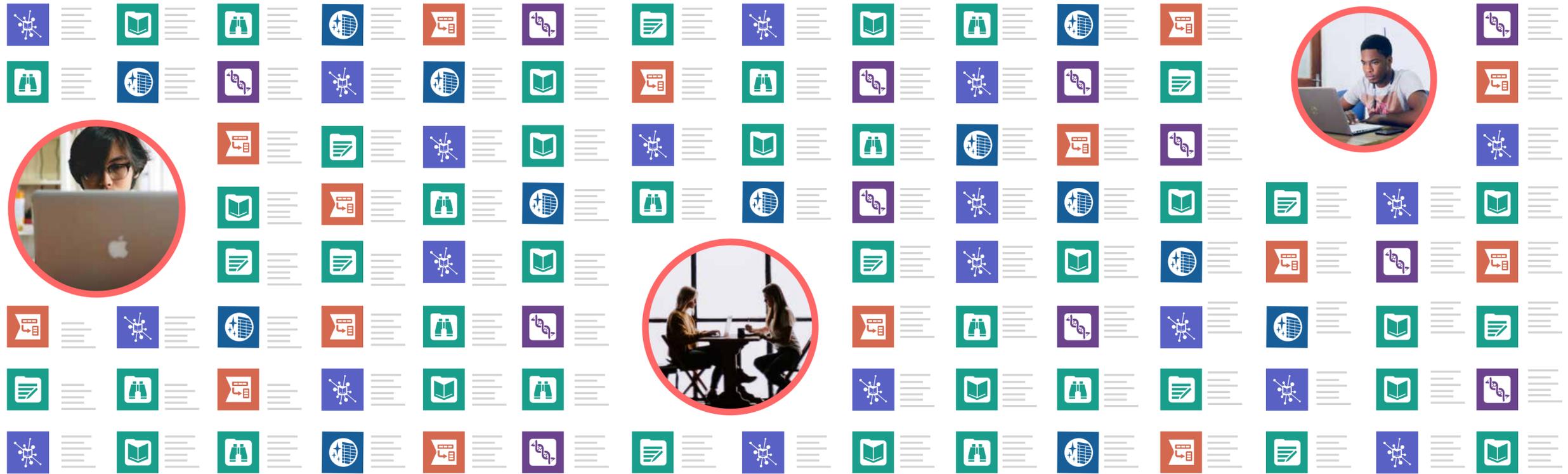


alteryx

E-BOOK



THE ESSENTIAL GUIDE TO

# DATA PREPARATION

From exploring to wrangling,  
prep your way to better insights



“ **6 BILLION HOURS**

per year are spent working spreadsheets

**26 HOURS**

are wasted in spreadsheets per week

**8 HOURS**

per week are spent repeating the same data tasks

”

— “The State of Self-Service Data Preparation and Analysis Using Spreadsheets,” IDC





CONTENTS

# LOOKING FOR A SMARTER WAY TO DO DATA PREP?

Data preparation in most organizations is time-intensive and repetitive, leaving precious little time for analysis. But there's a way to get to better insights, faster. We'll walk you through it step by step.

3

**Setting Up for Success:**  
What's the Big Deal About Data Prep?

4

**From Exploring to Wrangling:**  
The Major Steps of Data Prep

5

**DATA EXPLORATION**  
Discover Surprises

6

**DATA CLEANSING**  
Eliminate Errors

7

**DATA BLENDING**  
Join Datasets

8

**DATA PROFILING**  
Spot Bad Data

9

**ETL (EXTRACT-  
TRANSFORM-LOAD)**  
Aggregate Data

10

**DATA WRANGLING**  
Make Data Digestible

11

**Faster, Smarter Insights:**  
The Case for Analytic Process Automation

12

**The Alteryx APA Platform:**  
Why Use Alteryx for Data Prep?

# WHAT'S THE BIG DEAL ABOUT DATA PREPARATION?

The big deal is that you can't succeed without it. And that's not an overstatement. Data prep might 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 winds up living in silos, where it can't fulfill its potential, and in spreadsheets, where it's manipulated by hand. Silos and manual preparation processes are like a ten-mile obstacle course lying between you and the insights that should be driving your business.

If your organization is struggling with this lag time, you're in good company, as **69%** of businesses say they aren't yet data-driven<sup>1</sup> — but having other people with you in a sinking boat doesn't make it more fun to drown.

The more data you acquire and the more complex it gets, the more these problems amplify, so you need better solutions. What if you could work with any data format that struck your fancy? **WHAT IF YOU COULD AUTOMATE SOME OF THESE PROCESSES AND MAKE THEM FAST, TRANSPARENT, AND REPEATABLE?**

**IT WOULD PROBABLY BE A BIG DEAL.**

<sup>1</sup> Big Data and AI Executive Survey," NewVantage Partners, 2019.

# HERE'S WHAT A GOOD DATA PREP STRATEGY LOOKS LIKE.

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 EXPLORATION.**

Discover what surprises the dataset holds.

## **DATA CLEANSING.**

Eliminate the dupes, errors, and irrelevancies that muddy the waters.

## **DATA BLENDING.**

Join multiple datasets and reveal new truths.

## **DATA PROFILING.**

Spot poor-quality data before it poisons your results.

## **ETL (EXTRACT-TRANSFORM-LOAD).**

Aggregate data from diverse sources.

## **DATA WRANGLING.**

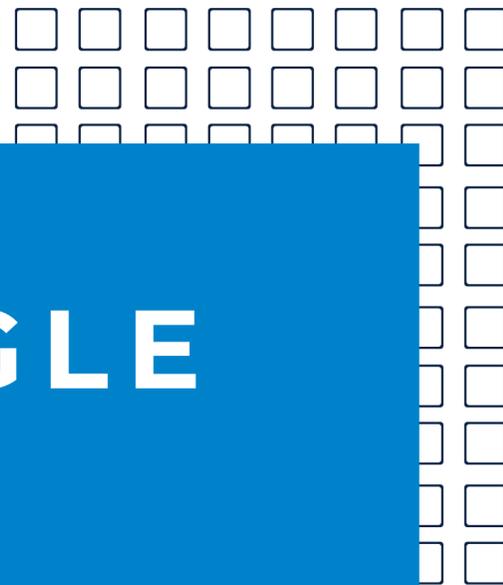
Make data digestible for your analytical models.

Ideally, as you're moving in and out of these activities, you want to record both your data and your process so that any mistakes you make aren't permanent, and so that others can repeat your results on their own.

Transparency and repeatability are the holy grail of data prep, but you can't have either in a spreadsheet-based system.

# IT'S A JUNGLE IN THERE.

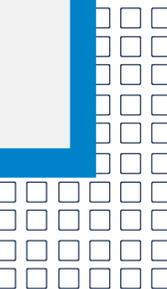
**BEFORE** you start working intensely with a new dataset, it's a good idea to step boldly into the raw material and do a bit of exploring. Although you might start with a mental picture of what you're looking for, or a question you'd like to see answered, it's best to keep an open mind and let the data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



## Here are some data exploration techniques that may spark big insights:

- ✓ Scan column names and field descriptions to see if any anomalies stand out, or 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 — say, 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 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.



# JUST SAY NO TO DIRTY DATA.

**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.



How to Clean Data Without Manual Edits

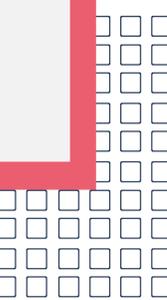


To prevent catastrophic losses like that, it's critical to scrub your dataset until it sparkles. As an analyst, you know this all too well, since this is probably how you spend most of your work week.

**Depending on the type of analysis you're doing, you need to accomplish six things in the cleansing stage:**

- ✓ Ditch all 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 info.
- ✓ Nip any unwanted outliers you discovered during data exploration.
- ✓ 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.

All these processes can be done manually, but at the considerable cost of your thinking time. Automated data cleansing tools, on the other hand, can do most of this work with a few quick clicks.

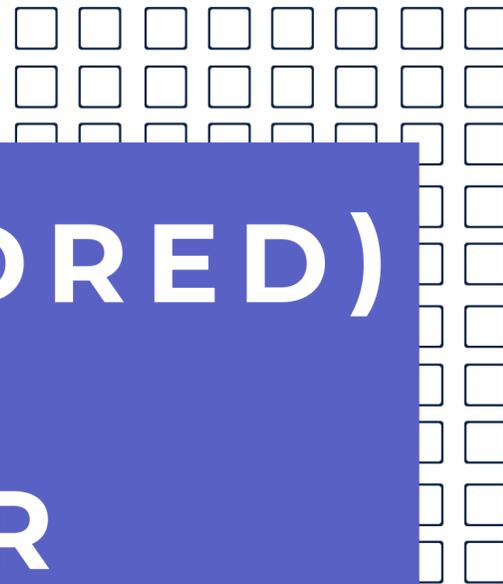


# TWO (HUNDRED) DATASETS ARE BETTER THAN ONE.

**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.



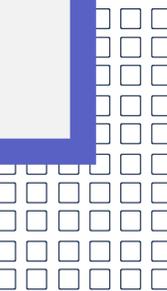
How to Quickly Input Data Sources in Alteryx



## 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. Think files and documents, cloud platforms, PDFs, text files, RPA bots, and application assets like ERP, CRM, ITSM, and more.
- ✓ **Blend.** In spreadsheets, this is where you flex your VLOOKUP muscles. (They do get tired, though, don't they?) If you're using self-service analytics instead, this process is just 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.

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.



# NOT ALL DATA MAKES THE CUT.

**DATA** profiling is a lot like data exploration, but with a more intense focus. Data exploration is an open inquiry performed on a new dataset; data profiling 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 that could have serious financial consequences for your company.

**There are three main 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.** What information does the data contain? Are there gaps or errors? This is the stage where you'll run summary statistics on numerical fields; check for nulls, blanks, and unique values; and look for 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? Examples might be formulas that connect cells, or tables that collect information regularly from external sources. Identify and describe all relationships, and make sure you preserve them if you move the data to a new destination.

Data profiling can be intricate and time-consuming. For a business end user to do it properly without the help of a specialist, data profiling software is a must.

# GET YOUR DATA DUCKS IN A ROW.

**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, which stands for "Extract/Transform/Load," and it's the centerpiece of a modern data strategy. ETL can also help you migrate data during a disruption — say, an upgrade to a new system, or a merger with another business.

## The three steps in a nutshell:

- ✓ **Extract.** Pull any and all data — structured or unstructured, one source or many — and validate its quality. (Be extra thorough if you're pulling from legacy systems or external sources.)
- ✓ **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.

The idea is to integrate all data and make it accessible to more people, not replicate the silos where it used to live. Forward-thinking companies look at ETL as a way to allow analysts, data scientists, line-of-business leaders, and executives to make decisions from the same playbook.

# ARE WE WRANGLING YET?

**THE** term “data wrangling” is often used loosely to mean “data preparation,” but it actually refers to the preparation that occurs during the process of analysis and building predictive models. Even if you prepped your data well early on, once you get to analysis, you’ll likely have to wrangle (or “munge”) it to make sure your model will consume it — not spit it back out.



## Here’s how you wrangle:

- ✓ **Explore.** If your model doesn’t perform the way you thought it would, it’s time to dive back into the data to look for the reason.
- ✓ **Transform.** You should structure your data from the beginning with your model in mind. If your dataset’s orientation needs to pivot to provide the output you’re looking for, you’ll need to spend some time manipulating it. (Automated analytics software can do this in one step.)
- ✓ **Cleanse.** Correct errors and remove duplicates.
- ✓ **Enrich.** Add more sources, such as authoritative third-party data.
- ✓ **Store.** Wrangling is hard work. Preserve your processes so they can be reproduced in the future.

Data wrangling is usually performed with programs and languages such as SQL, R, and Python. That requires technical know-how that the average analyst doesn’t have. To make this process accessible for your whole organization, you’ll need to use automated analytics software.



# DATA, MEET THE 21ST CENTURY.

What happens in a world 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?

**In our experience, automating data preparation looks like this:**

## **QUICK WINS.**

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

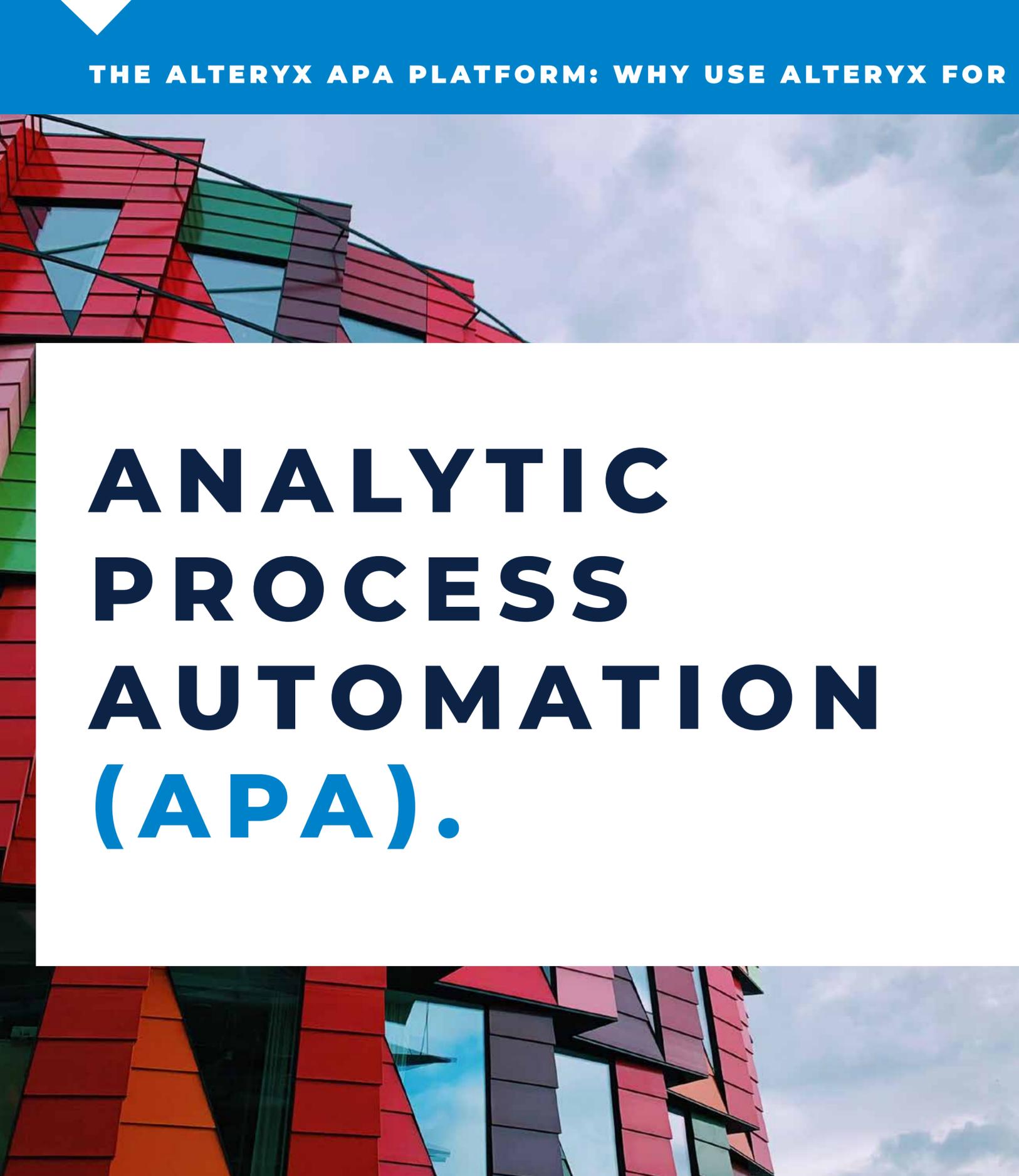
## **TIME FOR INSIGHTS.**

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.

## **PERPETUAL 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.

It's such a profound change — a different universe, really — that we have a name for it: **Analytic Process Automation (APA).**



# ANALYTIC PROCESS AUTOMATION (APA).

And your organizational ROI? Glad you asked.

1. TOP-LINE GROWTH
2. BOTTOM-LINE SAVINGS
3. DRAMATIC EFFICIENCY GAINS
4. FAST UPSKILLING OF WORKFORCES
5. RISK MITIGATION

---

*“We use [APA] across many of our businesses to leverage data, automate processes, and empower our people to become self-service digital workers.”*

— Rod Bates, Vice President, Decision Science and Data Strategy,  
The Coca-Cola Company

# START ANYWHERE. SOLVE EVERYTHING.

**Alteryx** is the only quick-to-implement, end-to-end data analytics platform that allows you — and everyone you work with — to solve business problems faster than you ever thought possible.

If you want **Analytic Process Automation** (and believe us, you do), we rock that space like nobody else. Our platform can 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 (or advanced modeling with code for all you data scientist superstars)
- ✓ Support for nearly every data source and visualization tool out there
- ✓ Performance, security, collaboration, and governance (translation: It's like sending warm brownies to your IT department)
- ✓ ROI and then some

## THE ALTERYX EFFECT:

Shrinking process times, speeding up insights, and generally saving the day.



Over **2,000 HOURS** saved  
in manual effort

— The Salvation Army

**1 YEAR** of store sales data  
organized in **1 HOUR**

— 7-Eleven

**\$80,000** saved annually  
using automation

— Amway

Time for analytics shrunk from  
“previously impossible” to **20 SECONDS**

— Chick-fil-A

**69%** **\$6M+**

**FASTER TIME  
TO INSIGHT**

**HIGHER ANNUAL  
REVENUE PER  
100 ANALYSTS  
EMPLOYED**



I simply can't do my job without Alteryx, nor would I want to.

— JAY CAPLAN

Senior Business Analytics Manager  
Coca-Cola

COCA-COLA



I built a workflow in **10 minutes** on our first day that queried **five billion** records in **20 seconds**. And immediately I realized there is something going on here that is really cool and powerful.”

— JUSTIN WINTER

Manager of Supply Chain Analytics  
Chick-fil-A

CHICK-FIL-A



Alteryx empowers people like us, who have little to no computer coding background, to do complex things with data even though we have no one in IT who can write Python. It allows us to follow the ideas in our minds and move from question to answer a lot faster.

— ALEXANDRA MANNERINGS

Director of Analytics  
Colorado Hospital Association

COLORADO HOSPITAL ASSOCIATION



It's crazy that we used to spend about **80%** of our time bookkeeping and **20%** of our time engaging with customers. But now with Alteryx, we've reversed that and have become an 80% customer advisory company with just 20% spent on bookkeeping. Through this process, it has enabled us to give our customers better experiences.

— BRIAN MILRINE

Business Strategy Director  
Brookson

BROOKSON



Alteryx pushes our analytics from playing checkers to playing chess.

— WILLIAM MCBRIDE

Director of Financial Planning & Analysis  
Cetera Financial Group

CETERA FINANCIAL GROUP



# DIVE DEEPER INTO DATA PREP

## JUMP FROM PREP TO INSIGHTS.



Check out the No-Sweat Guide to Advanced Analytics Success

## GET A TASTE OF DRAG- AND-DROP ANALYTICS



Try the Alteryx Data Blending Starter Kit

### ABOUT ALTERYX

As a global leader in analytic process automation (APA), Alteryx unifies analytics, data science and business process automation in one, end-to-end platform to accelerate digital transformation. Organizations of all sizes, all over the world, rely on the Alteryx Analytic Process Automation Platform to deliver high-impact business outcomes and the rapid upskilling of their modern workforce.

3345 Michelson Dr., Ste. 400,  
Irvine, CA 92612  
+1 888 836 4274  
www.alteryx.com

Alteryx is a registered  
trademark of Alteryx, Inc.

Gartner and the Gartner Magic  
Quadrant are registered trademarks  
of Gartner, Inc.

