Data Onboarding: A Survivor's Guide To Combining Unfamiliar, Disparate Data

Data Onboarding: A Brief Explanation

You may be unfamiliar with the vocabulary, but you know the feeling: you've got to deliver some analysis next week, and you can't even start the analysis until all your data sources are joined and standardized into one set. The engineering team can't spare a pair of hands, so it's on you to not only analyze the data, but do all the background work to organize it as well. Data onboarding—the preparation of unfamiliar data from disparate sources, both internal and external to the organization, usually without engineering support—is a serious challenge facing most business analysts.

Existing Approaches to Data Onboarding Are Inadequate

Great analysis can't be drawn from poorly-structured data. And yet, onboarding data from outside vendors or disparate data internally can be a Catch-22. Either it can be done right, but slowly; or it can be done fast and with problems. The first option involves asking engineering or IT for help creating complex technical solutions. Assuming these resources are available, the process is cumbersome and expensive, and it rarely happens quickly enough for a business unit to make timely decisions. The second option is to do it yourself, hacking your way to a solution using tools such as Excel, Access, or whatever tool you know best. This might get the job done quickly, but maintaining these fixes can be challenging, and the tools' various limitations make these approaches hard to share and scale across global organizations while maintaining data integrity and consistency.

The risks to the business are high because the most common outcome in both situations is that data issues surface during the process, which could delay the entire process. This can cause the project to be delivered late, leading to loss of business opportunity, or the customer/stakeholder to lose trust in the team due to slow or inaccurate results—or both.

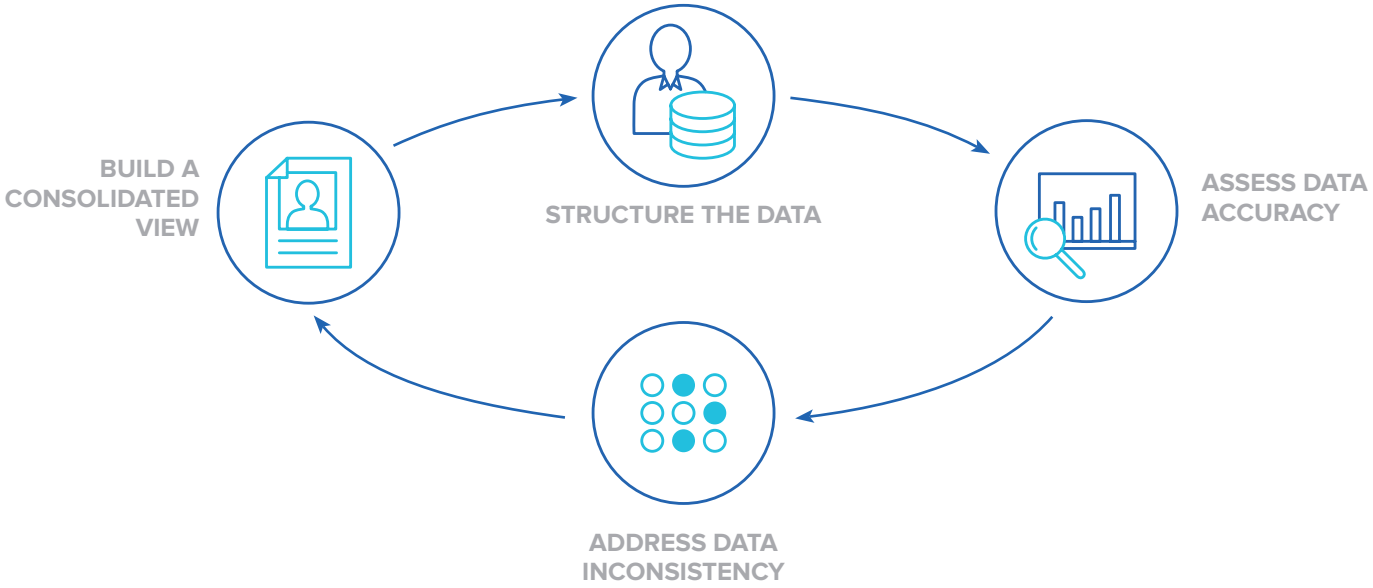
Frustrated? You aren't alone. Research shows that most analysts spend over 80% of their time preparing data for analysis, and data onboarding can be a big part of that, depending on your particular analytic focus.

Onboard Data The Right Way, From Day One

When it comes to data quality, every analyst knows it's "garbage in, garbage out." If data is being onboarded from uncurated or inconsistent sources, the analysis will be useless. Even when working with unfamiliar data sets and no engineering support, the goal of data onboarding should always be the same: create a well-organized, repeatable, and trustworthy process to generate accurate data. That kind of process requires more advanced technology than Excel or Access.

As the leader in data preparation, Trifacta's data wrangling technology can expedite the data onboarding process. With a visual interface powered by machine learning, Trifacta guides users through the process of wrangling data and accelerates the challenge of standardizing disparate data.

Ultimately, Trifacta's unique approach to wrangling data transforms the entire data onboarding process to exemplify the iterative nature of data onboarding, instead of the largely trial-and-error based process under legacy tools. Cleaning and joining data from disparate sources is an iterative process, starting with surface data issues like data structure, consistency, and formatting, to the more complex issues of duplicate and non-matching data. Here are the steps of the process and how Trifacta enables them:



STEP 1: STRUCTURE THE DATA:
Some data, such as APIs, JSON files or web logs, comes in complex hierarchy or in machine language and is therefore unreadable by humans. Any data like this must first be put into human readable form, such as a grid, so assessment can begin. Trifacta automatically recognizes data types and organizes them into a grid.

STEP 2: ASSESS DATA ACCURACY:
Are there missing values in cells? Maybe there are mismatched data values like an invalid zip code. You might even see value distribution anomalies, a form of unexpected data value (such as human life spans of 250 years, or percentages over 100 for fractions). Identifying a data set's flaws and reviewing for accuracy is critical to effective data onboarding, and it's at the core of Trifacta's architecture. Every time you open a data set or derive a new value from an existing data, Trifacta automatically and dynamically profiles your data, assesses its accuracy, and then displays a health check bar with accuracy information for each column.

STEP 3: ADDRESS DATA INCONSISTENCY:
Once the data has been assessed, and its flaws identified, the problems can be addressed. It's no surprise that this is the bulk of the data onboarding process. Starting with missing and mismatched attributes as well as anomalies and outliers, the data is then formatted and standardized to an organization's unique use case. Sometimes cross reference and lookup tables or conversion formulas are useful depending on an organization's needs. Lastly, duplicate and inconsistent records are addressed. Trifacta offers powerful tools to expedite, automate, and share this process to augment the common knowledge and team's efficiency.

STEP 4: BUILD A CONSOLIDATED VIEW:
Once the data is cleaned, Trifacta provides tools that Excel and Access can't in order to help you easily combine the data in a consistent view (and maintain it) offering an array of functions to join datasets, to pivot and unpivot data, aggregate values, and derive key indicators. If all goes well, it's on to building reports, creating your presentations and spreadsheets, or exporting into internal applications. If there's a problem, the process begins again until the data is finally ready.

Faster Data Onboarding = Faster Data ROI

Thanks to Trifacta, it's now possible to onboard data faster than ever, without compromising data integrity, scalability, or collaboration. In fact, the companies that have discovered how to solve this bottleneck are so excited about data onboarding improvements that they're going on the record to talk about it. Here are three companies who turned to Trifacta when their previous data onboarding process failed them.

PepsiCo: A 70% Reduction in End-To End Reporting Time

PepsiCo's sales forecasting team is responsible for collaborating with the world's largest retailers, each with their own data warehouse, to supply the precise amount of PepsiCo products to the right stores at the right time. Operating on razor-thin margins, PepsiCo relies on its sales forecasting team to churn out fast, accurate reports that blend retailer data with internal data. As many analyst teams do, they relied on manual Excel and Access for their data preparation but it was especially inefficient and error-prone. Using Trifacta, PepsiCo's team reduced end-to-end analysis time by 70%. Best of all, now that the actual report-building time is decreased by 90%, analysts can spend more of their time forecasting data. Millions of dollars in benefits were achieved in just the first few weeks.

"Trifacta brought an entirely new level of productivity to the way our analyst and IT teams explore diverse data and define analytic requirements. Our users can intuitively and collaboratively prepare the growing variety of data that makes up PepsiCo's analytic initiatives."

-Mike Riegling, Data Analyst, Pepsico

MarketShare: 10X Faster Customer Onboarding Process

As a marketing analytics company tasked with helping marketers grow revenue, MarketShare (now owned by Neustar) is on the front lines of data onboarding. They are helping marketing professionals adapt and manage a rapidly expanding portfolio of digital channels, each with their own data warehouse and structure. But to keep the promises they make to clients—a 3-50x return on the investment in their service in year one—they need to onboard their new clients' data quickly. But client data always came in inconsistent structures and formats—from Excel to text files, from APIs to JSON to name a few. Using Trifacta embedded in their internal application, MarketShare was able to experience a 10x improvement in the time required to onboard their new clients.

"[At] MarketShare. . .we have humans working with data to discover things and make judgments. In those cases where human intuition is critical to analysis, Trifacta is critical to us."

- Anna Dorofiyenko, VP of Data Science

MarketShare

NationBuilder: Aggregating Voter Data At Scale, In Near Real-Time

Some companies have a business model based entirely on the quality of their data sets and their services in analyzing that data. The data itself is the asset, and it must be deep, clean, and accessible in order for the company to maximize its market value. NationBuilder aggregates voter registration data from over 3,000 counties in the United States, which is a challenging task alone given the large, inconsistent, and poorly-formatted nature of county voter databases. With millions of records on a state, county, and city level being updated constantly, NationBuilder needs to onboard data in a way that's fast, accurate, easy to maintain, and scalable.

By using Trifacta, Nationbuilder was able to dramatically reduce the time spent building a national voter file, while eliminating the need for custom data transformation tools, thereby empowering a much larger and less technical team to derive insights and value from its important data set.

"Updating our national voter file in advance of the 2016 election was a business imperative. Trifacta helped to make that possible by simplifying our data operation and providing dedicated customer support every step of the way."

-Gina Davis, VP of Professional Services, NationBuilder

Trifacta: Data Onboarding In Record Time

Any organization dealing with disparate external and internal datasets and limited developer time can now expedite their data onboarding process with Trifacta's built-in, scalable, collaborative tools—with increased data accuracy.

Designed from the ground up to be self-service, Trifacta's user-friendly interface can enable more people than ever to analyze data to meet organizations' goals faster. To learn more about wrangling data for data onboarding, download the free Principles of Data Wrangling eBook here.

Ready to get started? Try our free desktop product, Trifacta Wrangler, here.

Questions About Trifacta? Email us at team@trifacta.com.