Data Ladder
Get the most out of your data

White Paper

Why In-House Data Quality Projects Fail

68 Bridge, St. Suite 304
Suffield, CT 06708

+1 888-779-6578

Sales@DataLadder.com

www.DataLadder.com

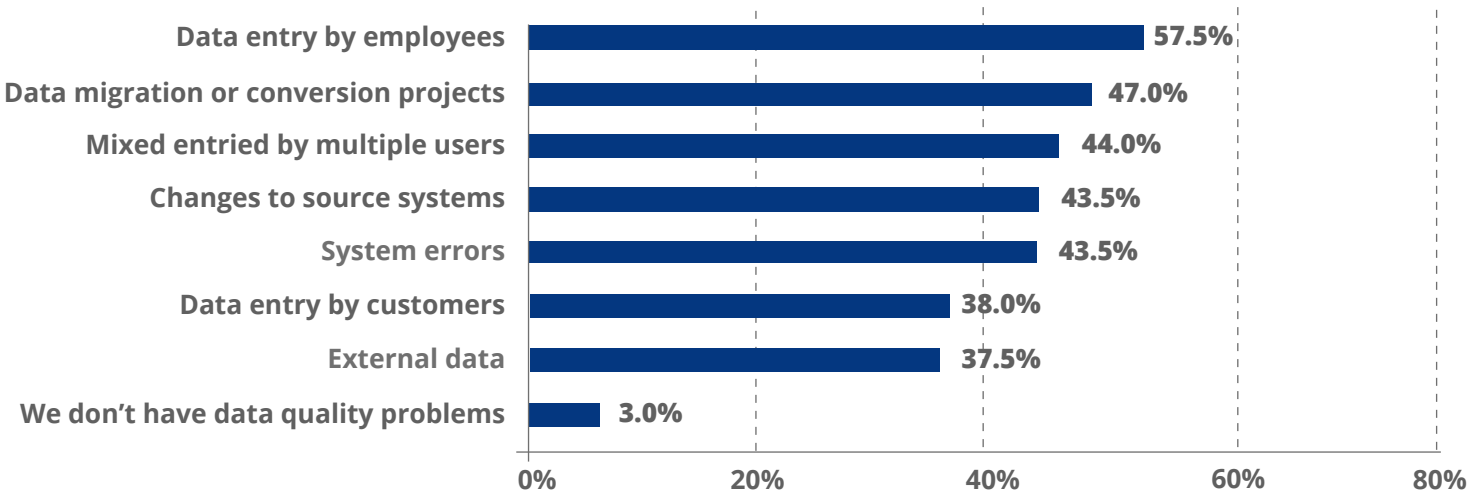
Data Ladder
Get the most out of your data

Why In-House Data Quality Solutions

Consistently Become a Liability for Businesses

THE STATE OF ENTERPRISE DATA QUALITY

Causes for poor Data Quality



Data Ladder
Get the most out of your data

Content

Introduction

The decision to build a software application in-house becomes almost a knee-jerk reaction when you have access to your organization's development team, even when research shows **that more than 70% of all in-house software development projects fail.** The arguments for doing so commonly include:

- Our developers can easily and cheaply build out a similar solution
- We will have direct control over future development and can scale it as needed
- No off-the-shelf solution provides the kind of specialized functionality we require
- We will have a complete understanding of how it works when we build it
- We will be able to develop precisely what we need
- We know our business needs better than anyone

Some of those arguments may have some merit depending on your case, but in over a decade of implementing data quality solutions for over 4,000 customers across 40+ countries, **we have consistently seen in-house data quality solutions ending up being a major liability for businesses.**

While the reasons for deciding to build an in-house data cleansing and matching solution are numerous, we can broadly categorize businesses taking this route into two major groups:

1. Those who minimize the importance of data quality and its impact on all business initiatives.

2. Those who think they can build out a better solution themselves, specific to their needs.

Type 1: Businesses that Minimize the Problem

We've seen that the top management in businesses that fall in the first category is typically averse to technology investments for something they don't consider business-critical, that is, data matching and cleansing. On the other hand, people in the same business who work with data down the organization hierarchy are all too aware of the repercussions of bad data (reduced productivity, lack of confidence in business data, tons of manual work cleansing data to be able to focus on their core tasks, etc.).

These are people who are familiar with the significant challenges and opportunities that data matching presents to the organization. They know the importance and the difficulties of doing it quickly, and even

But unless they are able to visualize the business value of data quality solutions, demonstrate it to management, and convince them to make the investment, these businesses usually continue to **spend time manually cleaning and matching data** and/or implement quick fixes like getting their developers to cobble together and store procedures in SQL that measure edit distance to match their data.

"How well an organization is run is a function of how well they manage their data"

Tip: If a business case is to be taken seriously, you must present it in the language of the business and speak to the critical and specific business priorities of key stakeholders. Read more on **How to Create a Business Case for Data Quality Improvement.**

What these businesses don't realize is the enormous financial impact that inaccurate and missed matches these workarounds have on their business in the long run. **According to Gartner,** poor data quality is a primary reason for 40% of all business initiatives failing to achieve their targeted benefits. How well an organization is run is a function of how well they manage their data.

See if your oganization is a Data Leader or a Data Laggard

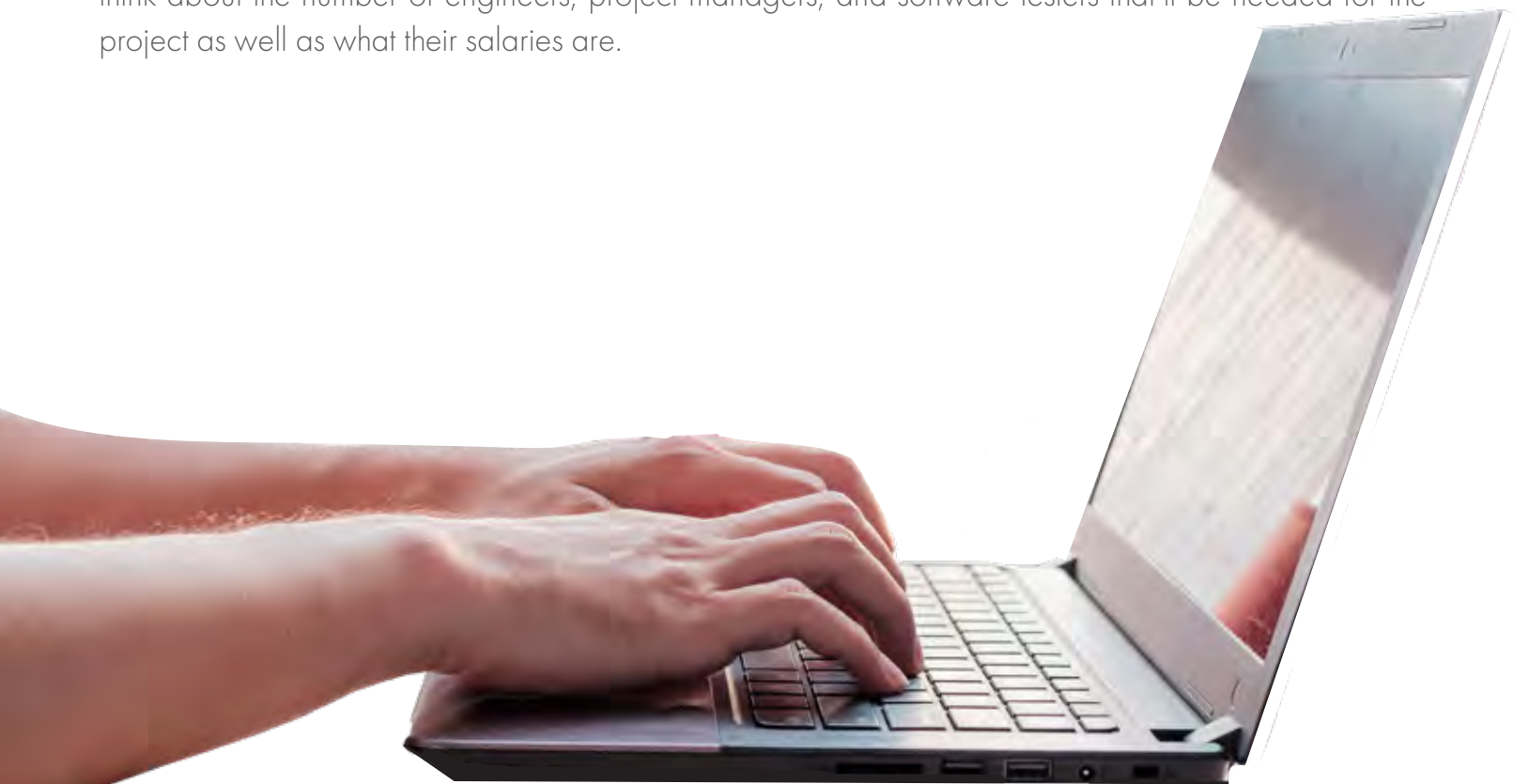
Type 2: Businesses that Think They Can Build a Better Solution

"When you begin a project, the software that you are 'going to build' always looks better than the software someone else already has because you haven't yet run into the limitations that inevitably show up in software engineering. As such, we will buy wherever we can,"

writes Timothy Campos, former Facebook CIO.

Among the top reasons businesses that we talk to decide to build their own solution is that they think they can clean and match their data better because they know the business best. They also perceive building an in-house as more or less "free" to build—you have to pay your engineers anyway, so it's a sunk cost—and free to deploy in perpetuity.

These businesses don't realize at the onset that software development and IT resources are still a cost: think about the number of engineers, project managers, and software testers that'll be needed for the project as well as what their salaries are.



Here's a quick cost analysis:

If you don't build it	
Annual cost of paying for a commercial solution	**$15,000**
If you do build it	
Number of employees required	**6**
Average employee salary + benefits	**$120,000**
Weeks to build	**20**
Days per month of maintenance	**3**

$275,975+	**$70,965**	**-$55,965**	**Never**
Cost to build	Annual Maintenance	Saved Annually	Years to Savings

And this is just the tip of the iceberg, as we will show you **further on in this white paper.**

As far as the belief goes that they'll be able to achieve higher match accuracy and that their in-house solution will be as flexible as they need it to be, these companies underestimate the complexity of their data quality problems greatly. From what we've seen in these cases, companies look at a few data sets during a point in time and build their solution around it, often not taking the time out to test their matching system extensively. Consequently, this leads to significantly lower match accuracy, missing as many as 1/3rd of matches.

Data matching isn't software development; the science of rule-based similarity is not the same skill as software development; so you can put together a team of the finest minds to build something that gets results, but it's not nearly the same. 13+ years of top tier domain expertise working with data matching scenarios from around the world, different data types, different use cases, different languages, etc., perfecting the workflow tools and our proprietary algorithms -- that's what enables Data Ladder to outperform comparable solutions from even IT giants like IBM and SAS in more than 15 independent studies.

Recently, while talking to a major industrial supplier, one of our solution architects was able to find nearly 30,000 matches in a data set with a total of 40,000 records, while the company was able to find just around 13,000 matches using their processes. Finding 42% fewer matches with their own solution, they were forced to make a significant compromise on the number of customers they were able to track for reporting and decision-making.

Another example is that of the **Preschool through Twenty and Workforce Information Network (P20 WIN)** program approved by the Connecticut State Department of Education and the Board of Regents for Higher Education. The board **published a report** that compares the results of Data Ladder's best-in-class data matching software with another advanced state-sponsored data matching solution. According to the report:

"In order to evaluate this tool, the State Department of Education (SDE) developed a file of student data that was matched to data from the National Student Clearinghouse (NSC) using NSC's proprietary algorithm, which resulted in 15,570 matches. This means that of the 38,426 students from the 2009-2010 four-year high school graduate cohort, the NSC has a record of 15,570 enrolling in a Connecticut public institution of higher education after high school graduation.

In contrast, the process of matching data through Data Ladder found 1,030 additional high school graduates enrolled in a Connecticut public postsecondary institution than were found by the NSC. Using the Data Ladder tool, analysts were able to review the output based on the thresholds for each matching criteria. Filtering the entire output data set for low threshold scores did not reveal any matches that looked incorrect upon visual inspection – producing an estimated 100% match rate."

Data Ladder
Get the most out of your data

Why In-House DQ Solutions Consistently Fail

In our decades of talking to organizations of all sizes, from Fortune 100 companies to public sector institutions in 40+ countries, we have consistently seen in-house solutions failing because of the following reasons, or a combination thereof:

Features of the Solution	Data Ladder	IBM Quality Stage	SAS Dataflux	In-House Solutions
Match Accuracy (Between 40K to 8M record samples)	**96%**	**91%**	**84%**	**65%-85%***
Software Speed	**Very Fast**	**Fast**	**Fast**	**Slow**
Purchasing / Licensing Costing	**80 to 95% Below Competition**	**$370K+**	**$220K+**	**$250K+**
Time to First Result	**15 Minutes**	**2 Months+**	**2 Months+**	**3 Months+**

Notes: The study compared 15 different products for the test above, using datasets from university, government, and private institutions (80K to 8M records). The results include the effect of false positives. To optimize speed and accuracy, multi-threaded, in-memory, NoSQL processing is recommended. Speed is important for accuracy because the more match iterations you can run, the more accurate your results will be. Benchmarked in 15 different studies, DataMatch Enterprise™ was proven to find, on average 5-12% more matches than leading software companies IBM and SAS.

Rated the fastest and most accurate data matching solution in more than 15 different **independent studies**, our solution enables clients to find 5-12% more matches than industry-leading solutions from IT giants like IBM and SAS while costing 90% less. As seen in the table above, in-house solutions scored significantly lower in this study done by the **Center for Data Linkage at Curtin University**.

Data Ladder
Get the most out of your data

"We saw a much higher proportion of matches within a 1-hour demo than what we've seen in our decade of matching."

Fortune 500 Manufacturer

On the surface it seems like building one's own matching solution is simple enough. Companies think they can simply use public matching algorithms to put together a matching solution. We've seen that in-house solutions typically incorporate single public algorithms, and offer a very cumbersome and simplistic approach. This reduces both speed and match accuracy greatly.

And, it's not just about the algorithms – it's more about the entire process flow, how well the process is managed end-to-end, how do multiple matching algorithms work together, which one takes precedence over the other and when, does the organization understand the issues in their data properly to be able to get the most matches, etc. From what we've seen, half the time, they don't even know what their data is with in-house solutions.

Here at Data Ladder, we go beyond matching. We focus on "data discovery".

By understanding the scope and nature of data problems in any project, provided to users in a concise profile view that pinpoints data issues that must be addressed to improve data quality, businesses will already have quantified the data issues that they would run into somewhere down the road while preparing their data, whether its for migration, business intelligence, master data management, standardization, or governance.

2. In-House Projects Have Too Many Dependencies

In-house IT projects depend heavily on the people who built. In most organizations, there's just one core member leading the project who knows exactly what's going on behind the scenes. For the rest, the solution is effectively a black box. No one else knows anything about how it works, the implementation, or how to fix any issues that arise.

So what happens when that core member from your IT team leaves? Often times, the project gets left to someone who doesn't understand it, or in many cases, the data matching project gets tabled. In the meantime, the allocated project budget still exists, so money is being wasted while the business need is not being fulfilled.

We've seen several cases where in-house solutions put the company at risk. One government institution needed to match between large populations for a study, so they decided to use a university to perform the data matching. Paying $500,000 a year to have it maintained, the people who created the code have since left the university, but now the institution is finding bad matches in the data, which now impacts a $10 million reporting project. This puts both the project and many careers at risk.

Many corporations have also experienced this issue, finding bad matches with internal solutions, but due to an internal employee leaving, the company ends up missing many opportunities.

Data Ladder
Get the most out of your data

The Consequences of Missed Matches

Match accuracy can have a major impact on any business. For, say a retail company, the difference between 80% match accuracy and 96% could mean a major reduction in the customers they're able to track to segment them better and therefore increase sales. Bottom-line: if you are missing matches, you're minimizing the impact on a lot of business initiatives.

Let's take a look at how missing matches impact businesses across a variety of industries:

- In healthcare, poor linkage can affect patient safety, treatment outcomes, and certain operational costs. **Studies show** that an average organization's duplicate rate is typically between 5–10% for a single hospital. Using their estimate of $50.00 per duplicate pair for an organization in hidden operational costs, a hospital that creates only 5 duplicates a day would end up spending $78,000 per year as a result of duplicates.
- In education, poor linkage can affect policy, funding, curriculum decisions, and even nearby real estate values. In the case of the Connecticut State Department of Education (**full example discussed above)**, each 'match' represents a high school graduate that enrolled in a postsecondary institution, and each missing match has a direct negative impact on improving education programs and workforce alignment in Connecticut.
- For Procurement departments, Tax departments, and people on the purchasing side, bad item matches result in millions of dollars in lost opportunities to reduce costs, hours, days, and weeks of manual work, and reduced insight into where the company is spending its money.
- For Researchers, Publishers, Data Brokers, and many others, bad matches results in bad information, poorly calculated conclusions, and other real-life impacts following the distribution of the data or the research.
- For e-commerce or other sales purposes, missed product matches results in suboptimal customer experiences, fewer upsells and cross-sells, tons of lost money, reduced market share, higher purchasing costs, lost customers, and disappointed shareholders.
- In law enforcement, poor linkage can affect staffing decisions, risk assessments, and community security.

On a more granular level, when matches are missed and business benefits are not realized from in-house data matching solutions, more often than not, the employee or the team responsible for building that solution takes the blame. We've seen many promising careers come to an abrupt end when a costly and time-intensive in-house software solution didn't pan out. As often as we see this, we also see people who understand the challenges and opportunities that data matching presents, convincing management on the right purchase. This in turn boosts their career by helping the company capitalize on matching opportunities to grow business and increase operational efficiency significantly.

By purchasing a solution that has proven its value commercially, offers excellent support, and has a variety of **case studies in your specific industry**, you can make the case to management without having to worry about the consequences you'd otherwise have to face if your team decided to build their own solution, investing hundreds of thousands of dollars and still missing a large percentage of matches (as seen in the examples above).

Conclusion

Data Ladder has worked with a wide variety of businesses that fall in both Type 1 and Type 2, as categorized in this white paper, both eventually choosing to switch to a commercial solution like ours once their in-house solutions became a liability instead of helping them improve their business. As shown earlier, more than 70% of in-house software development projects fail, and as far as data quality projects go, closer to 100% in our experience. Don't be part of the statistic - **our experts can help assess your data** quality challenges and provide a solution that best fits your needs.

If at present, you're managing your data quality through an in-house program, it's time to reevaluate how effective your program really is in terms of time savings, cost, and ease-of-use.

ABOUT US

Data Ladder is a data quality software company dedicated to helping business users get the most out of their data through data matching, profiling, deduplication, and enrichment tools. Whether it's matching millions of records through our fuzzy matching algorithms, or transforming complex product data through semantic technology, Data Ladder's data quality tools provide a superior level of service unmatched in the industry.

Why Data Ladder

It's simple: our user-friendly and powerful software helps business users across many industries manage their data more effectively and drive their bottom line. Our powerful software suite, DataMatch Enterprise, was proven to find approximately 5-12% more matches than leading software companies IBM and SAS in 15 different studies.

Free Download