cluedin

white paper

Where do I fix my dirty data?

2019 | cluedin.com

cluedin

Where do I fix my dirty data?

There are a lot of decisions to make in our industry in data. If you look at the respooibsilties of the Chief Data Officer, one of the predominant ones is data quality. One of the questions we often hear when we are talking with companies is “Where do we clean the data?”. It is a sound question, because the answer is not simple nor is it a universal answer. I can summarise that a company has 3 choices. Firstly, to clean in the source systems, secondly in a universal and combined hub and finally, to not clean at all. Those that chose option 3 will not be around for too much longer.

"The CDO must determine the company’s current data quality and maturity levels – of which there are five. (1) Uncertainty, which typically involves the organization stumbling over data defects as programs crash and employees complain. There’s no proactive improvement process in place. (2) Awakening, during which a few individuals acknowledge the dirty data and try to incorporate quality in their projects before formal enterprise-wide support arrives. (3) Enlightenment is when the organization starts to address the root causes of dirty data through program edits and data quality training. A data quality group usually emerges here. (4) Wisdom arrives as the organization proactively works on preventing future data defects, and data quality incentives arrive. (5) Certainty emerges as the organization shifts to an optimization cycle – continuously monitoring and improving its data defect-prevention process.”

This quote from a prominent analyst firm is a nice look into what the market sees with the responsibilities for the CDO in the Data Quality field.

Let’s take the first option, to clean in the source. Admit it, it makes sense to clean data in the source and in fact many companies can and have had success with this. If we look at first principles, it makes complete sense to go back to the root of the problem and fix it there. Easier said than done, as believe it or not, the root is usually us i.e. humans. All the form validation in the world cannot save us from still entering bad data. The problem and flaw with this approach is that although data may become “cleaner” it does not conform to the values in other systems. To properly gain this view of the entire landscape you need to consolidate data into a consistent environment.

So then we explore the second option, which is to let bad things happen in the source and fix it in a central system which can then propagate changes back to the source. This is also quite flawed in nature. The benefits come in the form of “Clean once and benefit many”. The other benefit is consistency i.e. we standardise at a central level on the semantics and then source systems are cleansed periodically. Some would argue that there is the need for a central governance server where all systems can validate against that - I am not aware of a product that does this. The problem with this approach is that the data entry people don’t learn - they continue to put bada data through the source.

It really comes down to priorities. The cleaning in the source will definitely cost more money and will raise the quality of data, but not necessarily the standardisation, especially in large companies. The second option is the better of the two, but still not perfect.

So what should be the answer?

I am afraid I come with bad news. The right answer does not exist yet because the data quality industry is still very immature and non standardised. Here is what is necessary. There are many pieces to this. The first starts with realising that there is historical data, data for the future and unknown. With this mix, we can start attacking the problem at many angles.

The first is to patch what is broken.

For this we need a centralised governance system that a business can gather all the questions they need to answer and answer them in a central place e.g. Across the board, are we using a Universal Date Format, what format should phone numbers be, how do we universally represent genders. From this, the source systems should take their validation. In this way, if things change in the future, we have the central brain that makes sure the changes are instant and updates are handled.

The second is historical data.

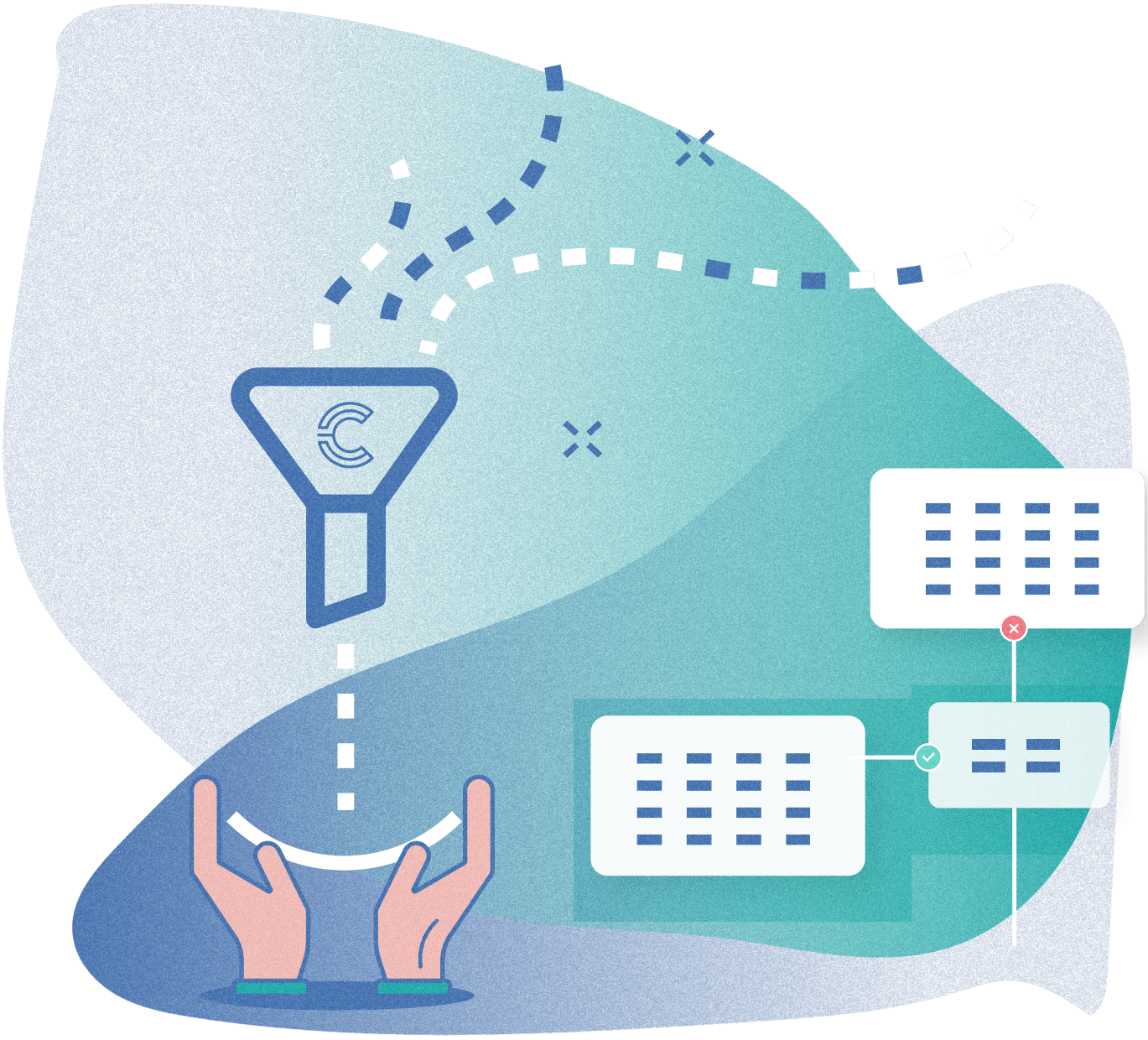
Without a doubt, the second option (data hub) mentioned above is the way to go. For one simple reason. You need to understand your entire data world and landscape before you can standardise. You also need to know the business rules to be able to standardise e.g. how often do you find that there is a very good reason why one system maintains a non standardised version of data, it happens all the time. This leads to one truth. Source systems will never be standardised. Hence the idea of layers on top of source systems is without doubt, necessary. It is important however that these rules are registered so that users are aware of why it is so. A central hub can always project out a standardised value to upstream consumers.

We will also often find that companies will take a system like the CRM and will spot enrich and clean it. It is important that this data is proliferated throughout the rest of the business (without breaking terms and conditions of the data cleaning provider). We have seen this action come when budgets are dedicated at a department level and one department prioritises it and others do not. Data cleaning and Data Quality should not fall under a department budget, but rather a global company budget.

There is a final option I will touch on here, which is the wrong option by default, but many companies are taking this approach.

The approach I talk of is cleaning data during the ETL process. “Wrong” is most likely the wrong word, because you can clean quite a lot of data using standard rules e.g. “Trim Whitespace” and “Does this match an email pattern”. To some degree, this is not the problem that companies are mostly struggling with. The problem is that in such a disconnected data world, with such an array of different tooling, that we need to solve the problem about the data that does not get cleaned via this standard cleaning. I refer to, of course, fuzzy data. Fuzzy data is the reason why ETL gets us only “so far” but for some “far enough”. Fuzzy data is the reason why most companies don’t have a GDPR solution. Fuzzy data is why companies worry about what high risk data they have in file shares, emails, document repositories and more.

The important part is that you establish a plan for how to clean your data. Goals and SLA's should be put in place to make sure that the levels of quality aware raised over time. This can only happen if you have established processes that measure, benchmark and improve quality over time. CluedIn addresses this with two main data steward tools. The first is **CluedIn Clean**, which is targeted towards allowing data engineers to improve and normalise the data. The second is a tool called **CluedIn Train**, which is targeted toward involving the business is labelling and curating the data.



Now that we have established some plans around where to clean data, the next important question to ask is “who should clean the data?”.

The answer is not as easy as saying “the person or persons who caused the dirty data in the first place”. The answer is not easy because it is not always apparent to all people what is “dirty data”. Still to this day, most end business users are not aware that using open ended text fields that allow you to put any content will cause unstructured data issues. Form validation can of course help and more structured fields can help as well, but this does not mean that unstructured fields don’t have their place i.e. notes and comments. The truth is that many people will play a role in cleaning data. Even if data is created perfectly in the source tool, that doesn’t mean it is normalised and clean. You can probably imagine that different tools will validate data differently and hence the need for centralised cleaning will always be necessary. The devil is in the detail and hence business users that are “close” to the data should play a role in supplying “context” and data engineers should be there to supply and apply normalisation.

How realistic is that business users will play a role in data cleaning? Some would argue that if we are not putting clean data into a system, then why would the same person clean this data at a later point? The reality is that business users should not “correct” data in systems like CluedIn, but rather let a system like CluedIn know that the data is simply incorrect. It is this balance that allows systems like CluedIn to “learn” over time of what looks like dirty data and what is clean.

In summary, it is obvious that cleaning data is on the highest of agenda’s for most companies today. It is the roadblock that is stopping value being gained from Business Intelligence, Machine Learning and more. Make sure that if you are charged with making your company data driven, that you make sure that delivering clean data to your business is a key part of the strategy.

cluedin