CluedIn

white paper

Solving the data trust trilemma

www.cluedin.com

Solving the data trust trilemma

Overview

There are many factors upon which decisions can be made. Gut instinct, personal preference, past experience, peer recommendations, well-established facts and of course, data. If the past decade has shown us anything, it is that organisations are increasingly turning to data to inform their decision making. This can be in any number of ways – in order to analyze customer behavior, monitor the activities of competitors, spot emerging trends and identify inefficiencies and wastage. Once the data has been gathered and interpreted, that intelligence can then be used to make decisions, and in turn to assess whether the expected outcome was achieved.

> *"Whenever you see a successful business, someone once made a courageous decision."*
>
> **Peter F. Drucker**

Increasingly, data is also being used to fuel projects like Machine Learning (ML) and Artificial Intelligence (AI). ML projects need huge data sets to be effective, regardless of whether this is supervised, unsupervised or reinforcement learning. Similarly, in order for an AI algorithm to output any prediction, it has to be fed with large volumes of data. This is likely to be qualitative and quantitative, and could be structured, unstructured or semi-structured.

There are numerous examples of companies that are using data in this way – such as Google, Facebook, LinkedIn and Netflix. These organisations are pioneers, and can genuinely be considered to be "data-driven". The majority of companies however are still struggling to put their data to work in a way that demonstrably benefits the business. In this paper we discuss the number one reason why this is so challenging – a lack of trust in data on behalf of the business – and how to solve it.

The missing ingredient

With so many decisions and the success of critical business initiatives dependent upon data, it is no wonder that companies are investing heavily in data and solutions to manage that data.

According to Grand View Research, the enterprise data management market accounted for USD 61.95 billion in 2019 and is expected to reach USD 135.88 billion by 2027.

> And yet, despite massive investments in technology – such as Data Lakes and modern Data Warehouses – and even after hiring large teams of Data Scientists and Engineers, many enterprises still feel as though something is missing. As if they are not really data-driven at all.

This is because there is something missing. That thing, in the majority of cases, is a reliable supply of trusted data. Without the right data foundation in place, it is close to impossible to provide valuable data in a meaningful way. The result is poor decisions with disappointing or no business impact, or no decisions at all. In the case of ML, the old adage "garbage-in, garbage-out" carries a double warning - first in the historical data used to train the predictive model and second in the new data used by that model to make future decisions.

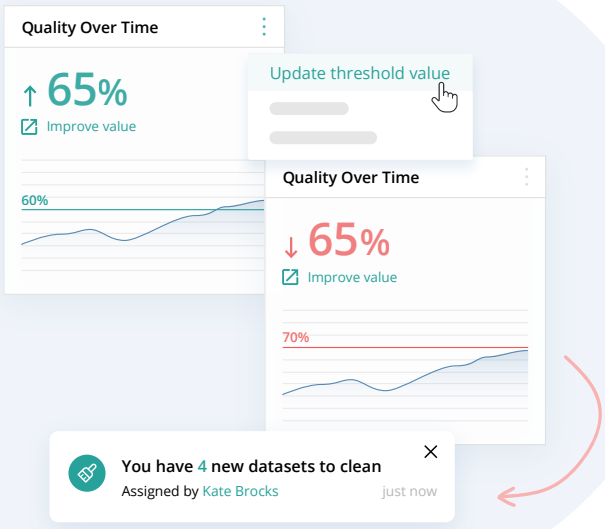
> *"If Your Data Is Bad, Your Machine Learning Tools Are Useless".*
>
> **Harvard Business Review**

When data-driven projects go wrong, it is natural to blame it on bad data. We would argue that there's no such thing as bad data. What that really means is that the data has not been given the right amount of TLC to be useful. The end result though is the same – a misalignment or failure to achieve the desired business outcome, and a loss of trust in the data and the projects it is used for.

The strongest relationships are built on trust, and it is no different when it comes to data and the business. When trust in data is lost, it is hard to win the support of the business for future data-driven projects. Fixing this requires attention in three key areas, and a transparent way of showing the business that your data has the required integrity and consistency to be believed.

What is the data trust trilemma?

Trust in data comes from three areas - quality, lineage and governance. Let's examine each one in turn.



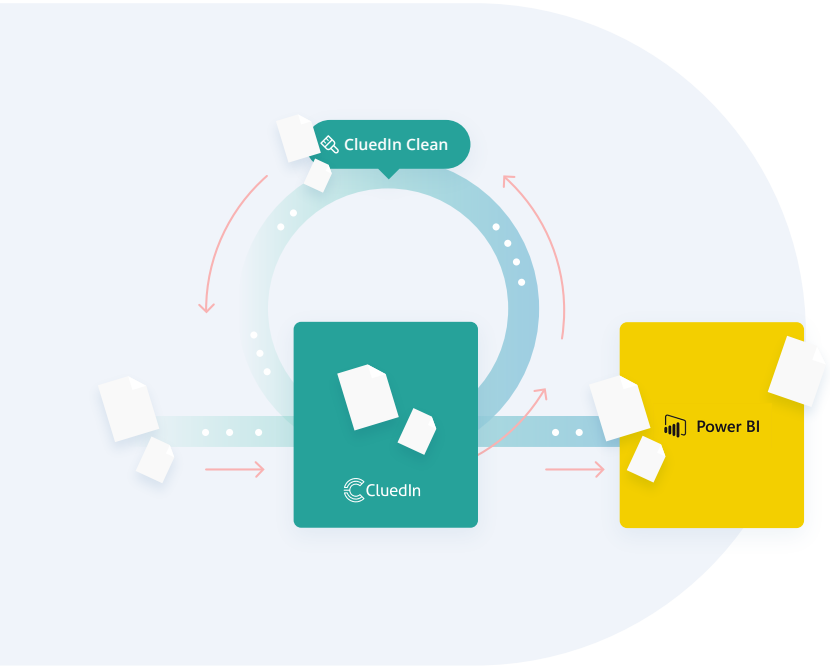
Data Quality

It might seem obvious, but poor quality data is the number one enemy of data-driven decision making and data science initiatives. The scary thing is that most organisations have no idea how good, bad or ugly their data is. If you have no idea of the state of your data, you stand little chance of fixing it. One of the quickest and most effective ways of winning back trust in your data is to be completely honest about the quality of your data. Even if it is terrible.

Of course data quality is subjective and can mean different things to different people. Which is why it is so important to agree on the metrics you will use to measure data quality. There are lots of data quality metrics you could choose – accuracy, relevance, stewardship, consistency and uniformity to name just a few. The key is to decide on which are the most important to you and measure against them over time, clearly demonstrating improvements.

It is also worth remembering that 100% data quality isn't a realistic goal, simply because data changes all the time. But if you started with an overall data quality of score of 34%, and in a matter of weeks have managed to drive that up to 76%, you are well on your way to proving to the business that the data is of a sufficient quality to be trusted.

> *"On an important decision one rarely has 100% of the information needed for a good decision no matter how much one spends or how long one waits. And, if one waits too long, he has a different problem and has to start all over."*
>
> **Robert K. Greenleaf**



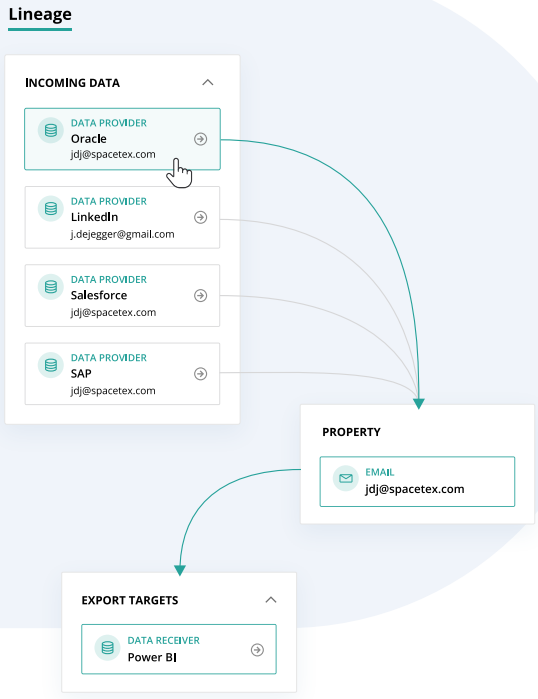
Data Governance

At its core, data governance is about managing data throughout its lifecycle – from acquisition, to use and disposal. It includes the technologies and process you put in place to make sure that your data is secure, protected, available and usable to the right people at the right time, and that you have visibility into where that data resides and its classification at all times.

Data governance goes hand-in-hand with data quality and you can't have one without the other. Think of it as data governance providing the birds' eye view, as it focuses on the overall management and visibility of data, with data quality paying attention to the integrity and accuracy of the data itself.

When correctly implemented together, the benefits of data governance and data quality include:

- Making better, more well-informed decisions
- Managing costs more effectively
- Improving regulatory compliance
- Managing risk more easily
- Data democratization as more people have access to the right data



Data Lineage

Data lineage is the journey your data takes as it flows through your organisation. Very similar to an audit trail, data lineage allows you to identify:

1. The origination of the data
2. When the data became available to the business
3. What happened to the data as it moved from source to target
4. How the data got from source to target
5. Who was responsible for the various steps along the way

Data lineage is a facet of data governance, and an exceptionally important one when it comes to establishing trust. Imagine presenting a report in a management meeting and being asked where the data came from. Being able to explain how and when the data was acquired, how it has been stored, who has modified it, etc. would be extremely helpful to building confidence and trust in the recommendations you are making based on that data.

You can also imagine how critical full lineage is when responding to data protection requirements such as Data Subject Access Requests, for example.

The Holy Trinity of Trust in Data

These three disciplines – data quality, data governance and data lineage – are the pillars on which data trust is built. Each is closely affiliated with the other, and there are overlaps between them, but at present there is no single system which delivers the required depth of functionality to address the full spectrum of data trust requirements.

> Instead, organisations need to invest in a common data platform which can deliver all three as part of a set of fully integrated services. Your ideal data platform will also include provision for modern data warehousing, and analytics and Business Intelligence tools.

The main advantages of a consistent data ecosystem is that it helps to close the gap between databases and analytics products by delivering high quality, trusted data. It also allows you to achieve maximum agility when responding to shifting market conditions and customer demands. In addition, product and service innovation thrives in an environment where the necessary intelligence can be acted upon at speed.

Solving the data trust trilemma will only become more important as organisations continue to face unpredictable and changing macro and micro factors. Your data is the key to unlocking the potential of your business – and in many cases is the difference between success and failure.

CluedIn | www.cluedin.com