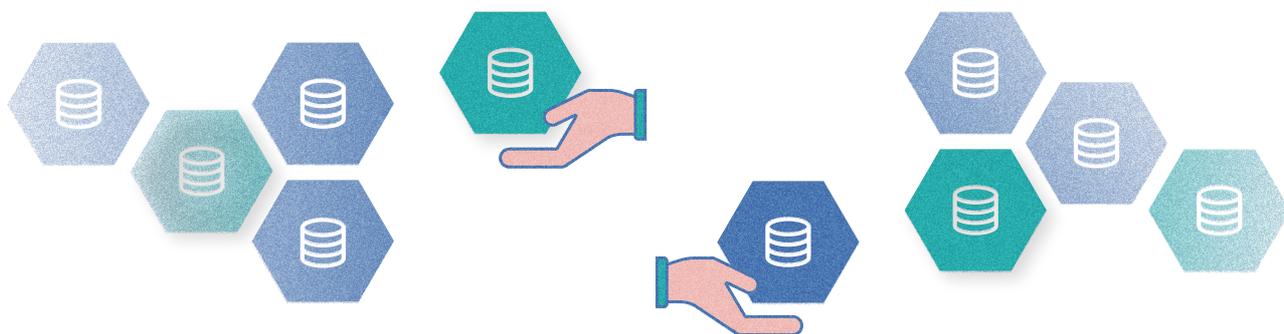


Why don't you have timely access to data?

Why don't you have timely access to data?

Let's admit it. You have all the technology in the world at your enterprise company, but the simple fact is that timely access to the data you want is not a reality. Why is this? You have a data warehouse, wasn't that supposed to give you this? Yes! In the old world. The datawarehouse was THE solution in the old world. We changed this world 10 years ago and it is only really starting to hurt companies now.

We need to break away from this idea that the Data Warehouse is your solution to this, but important to remember that it is an essential part of the solution. We also need to make sure that we don't just rely on the Data Lake as the solution. Once again, it is part of the solution. We also need to make sure that we don't make the Data Hub the solution (i.e. CluedIn) as it also is one part of the overall solution.



I would concur that the Data Lake definitely gives us timely access to data, but is it time well spent? I would argue that with the amount of work that is necessary to do after the lake, then timely is correct, but readiness is not. Ask yourself, do you want timely data or ready data. Both are useful for different use cases e.g. if my goal is to show as realtime as possible signal data from wind turbines, then you can argue that even the lake is not necessary but rather a live stream of data from source to a visualisation tool, but if I wanted a single view of my customer then we need to take the time to clean, blend, track the data and more. At CluedIn, we are not here to solve the first example of live data, we are here to focus on the pipeline needed to provide clean, tracked, blended, normalised data to the business and this is really where companies are struggling.

Let's jump back to the Data Warehouse for a moment.

I would argue that you have timely data to the Data Warehouse, but with the timeliness comes rigidity. So then why don't you have timely access to clean, accurate, governed, enriched data? Well the answer is that this has not existed for a very long time. The Data Hubs of the world are the first types of products to address this use case and to be honest, the field is new. Don't confuse the Data Hub with the Data Warehouse, they exist at different parts of the pipeline. It should not be a shock that this doesn't exist, it is the natural progression of technology that we solve things when they become a necessity, but now very much is the time for these types of technologies.

How can you test at work if you have this issue?

Simple. Ask your IT team for certain data, wait until you get it and then tell them that you need slightly different data than original requested. If their answer is it will take another few weeks, then you do not have timely access to data. There are other factors that play into this such as queues of requests and more, but the core issue still does spawn from inability to make data readily available to the business. Solutions like BI platforms will argue that they deliver realtime insight to your business. I agree. But we need to be more specific. It delivers real time aggregated data to the business. This doesn't help the Machine Learning team, this does not help the team that needs data to fuel a chatbot, this does not help a large percentage of the teams.



I often talk about the Data Hub as being a great solution, but what are the bad parts about it in respect to timely access to data. A generic data platform will never be optimised as much as it can for very specific use cases. For an example, imagine that you wanted to use your disparate data on your customers to fuel personalisation on the website. A Data Hub will get you the data but is it in the optimised format and storage, no, typically, never. However we should not fret. The good news is that we can use the Data Hub to prepare Data Marts that are optimised for the use case. Of course, this will cost you money to store the same data in different locations, but as we always say at CluedIn - when building a data pipeline you will typically have a few dials to turn, including cost, scale, performance, quality and more - as soon as you change one dial, the others move up and down in relation.

In summary, the reason why the Data Hub gives a good balance of timely, accurate, governed data is that in the large percentage of end use cases we are seeing today, there are some critical pillars that all use cases have in common i.e. we need blended, governed, clean, enriched data. The Data Lake is a little bit early for this maturity and the Data Warehouse is a little bit too late. The Data Hub is a nice balance of data maturity and performance that will help you cater for more diverse use cases.

