cluedin

white paper

What is the difference between the Data Warehouse and CluedIn?

2019 | cluedin.com

What is the difference between the Data Warehouse and CluedIn?

It is a relevant question. Very relevant. The Data Warehouse is the place that you place all your data in, to make a universal, catch all use case for any data project. Or so we thought. That is the thing with technology, it is hard to predict. Turns out, it was great for reporting. Turns out, it is still good for reporting. But also turns out that many use cases with data today can't use the Data Warehouse. Why? Because the data is too mature at that point. Sounds silly, but makes complete sense.

The Data Warehouse is designed to be able to run very well known queries on extremely large amounts of data. Guess what? That is not what is needed from the business today! We need flexible, easy access to data from across the business and the less time I spend in cleaning, blending and moulding it, the better.

But isn't this what the Data Lake promise? Easy access to data for the Data Science team.

Easy access, yes. Ready to use? No. You could argue we are a step in the right direction, but still we are not to the point where we can start to get value from data. We also can't forget that you still have to get data into the Data Lake! Much easier said than done. So as you may have already gathered, we might need something that is in-between the data lake and the data warehouse. The good thing is that our industry has already given this a name(s). The Data Hub or the Data Fabric.

The Data Hub is targeted towards being in between the Data Sources/Data Lake and the consumers of data. Why?

Because a large percentage of consumers of data need the same thing i.e. blended, clean, enriched, governed, tracked, normalised, catalogued data. This is most likely something that you are doing today for the Data Warehouse, but to achieve something more foundational and flexible, you need to abstract it a bit more. Let's then ask the critical questions and play Devil's Advocate.

If the data is too mature in the Data Warehouse, what level of maturity is the Data Hub and why is it at the right level of maturity?

The Data Hub is targeted at being generic, as generic as possible. This of course means that when you want to achieve something specific, that there will always be some extra work involved i.e. this is why the Data Warehouse is still needed in this picture. The aim of the Data Hub is to find the middle ground in not over-maturing the data, but maturing it enough to limit the work needed for the specific use cases. I can hear you yelling at this paper, thinking, yes that is what the Data Warehouse was supposed to do! Yes, I agree. But no-one had the use cases in mind for today's demands. Machine Learning, Data Science, Advanced Analytics, Flexible/Ad Hoc Business Intelligence, Personalisation, Data Privacy.

Can we all agree that in general, there are certain things we need to do with data to prepare it for use cases? Can we also agree that in general, dependant upon the use case we might have to do more preparation and others? Would we agree that the majority of cases require integration from many sources, cleaning, blending, governance, lineage? If the answer for you is "yes", how many of these use cases are you having success with when they data is coming from the Data Warehouse? Probably many. In fact, reporting and business intelligence are probably the parts that you answered yes to. How about every other use case where you need data? Not so much success right?

If we can now agree on this, we can also agree that a Data Hub should take care of the general "stuff" i.e. integration, cleaning, blending, governance, lineage, cataloguing and accessing this data easily. Welcome to CluedIn. CluedIn is exactly this. It isn't a Data Warehouse, it isn't a Data Warehouse replacement. It essentially does a lot of the preparation of the data for the Data Warehouse but only to the point that the other use cases can also use the same data. In saying this, the Data Warehouse would consume its input off the Data Hub.

Let me play Devil's Advocate again. But what about all the work we have already done in ETL to get data into our Data Warehouse, what about all the SSIS packages? Do we just throw that away? Not necessarily, but potentially some of it could be done better. This introduces friction into the situation, because suddenly that code we have in our ETL pipeline that has been there for 10 years is suddenly questioned and potentially obsolete. Yes. Yes it is. There becomes a point where if a business does not adapt to new ways, they fall behind. This is one of those examples.


cluedin