cluedin

white paper

What type of Data Catalog do you really want?

2019 | cluedin.com

cluedin

What type of Data Catalog do you really want?

The Data Catalog product of today is built with the idea of how we work with data from some time ago. The Data Catalog exists, because most companies have very siloed data. With the introduction of the Data Lake, nothing has changed, we still "need" a way to be able to know what tables, files, databases we have lying around our workplace and then we need to run the data preparation on them to realise their potential. This type of Data Catalog will soon no longer be needed as much as we need it today. Why? Because this way of treating data is the exact reason why we can't get value out of our data today.

Let's walk through the normal process of a Data Catalog and then highlight the wins and flaws of this approach.

Things usually start with the business requesting some data e.g. We need to know all our customers from across all locations. The IT team will then type the word "customers" into the Data Catalog and a list of results would come back with Tables, Databases, Applications, Columns with the name "customer" in it.

Here is the first fatal flaw. We are assuming that all systems call it "Customers", but there is no need for an assumption - they don't. Let's just assume this actually did work and that we have been a good company. Now we need to blend the data together. The IT team now needs to be business domain experts to figure out how to blend the different customer tables and databases together to be able to deliver the list of unified customers.

Flaw number 2, data doesn't blend naturally and it is close to impossible for the IT team to realise that they have to join the Customer Table, onto the Product table to be able to join back onto the customer table of another system. Let's continue, thinking that it was a perfect world and the data blended.

The next problem is that there is no science (yet) to what processing should be done in order to properly blend data i.e. should we clean before we merge? Or merge before we clean? Enrich after or before merging? The fact is that the ability for data to be prepared will change the ordering of this process every single time. It is really up to the data to determine what ordering the processing should be done to prepare data for the business.

So now that this business request has been fulfilled and now we can prepare a data mart in the catalog that is based off this request, so that when anyone else wants to use this exact dataset, then they can.

Would you rather have skipped many of the steps? Why would you want to do this more than once?

Well, if you are using the wrong technology - then you have no choice. But imagine you are using the right technology so that we only have to work with the business domain experts once, we only have to blend data once, we only have to clean data once. Pipedream? No, really it is not.

The data catalog of the future is enabling the ability for the business to search through pre-blended, clean, enriched data. Why? Because this is the solution to your data driven problems. It is the reason why business requests for data usually take 6 months or more.

However the Data Catalog we use today is still useful, but only for some.

Imagine you have taken the approach that because you have so much data, and hosting it in a Data Hub like CluedIn would be too expensive - that you want to only bring data in when you have a request for it. I think that could make sense. This is where the old type of data catalog makes sense, because it will allow you to easily find what data you need to bring into the Hub, but this type of Catalog should NEVER be exposed to anyone in the business.

However we do receive quite a lot of requests where the business would still like to know what data is available for them to consume. Great! This is where the new type of data catalog comes in, something that exposes what types, what quality, what properties are available to the business. An example would be something like “I want all our customers where we have 100% of the required data and they are in Denmark. “ Notice that at no point did I mention any tool. If we did, we would be thinking in a siloed manner i.e. “I need all our customers from Salesforce” - if you hear this type of thing then you need to be aware that you will never have the full view of your customer.

So we have established that the Data Catalog that you have potentially already invested in makes sense. Hopefully you have seen why it is more relevant for a different crowd i.e. the IT department that is in charge of delivering data to the business and in-turn delivering it to the business orientated Data Catalog.

cluedin