cluedin

white paper

What is connected data and why is no-one doing it?

2019 | cluedin.com

What is connected data and why is no-one doing it?

We are living in a relational data world. The sad thing is, that the majority of companies have data that is more like a Graph or Network in reality. Although this might be considered a moot example, there is a good reason why Google, LinkedIn and Facebook all fundamentally base themselves off the Graph structure - it is because it is the super structure, it is the structure that supports all sub structures. The trick that these 3 big players have identified is that the Graph is at the core, but it is supplemented and supported by other types of technology that can fill in for the parts where a Graph falls down.

Connected data is the idea that we can build up a network of discrete entities that are connected via relationships. Relationships are never always direct, they can be indirect - and if the intention is to find hints to relationships that could exists, then this will never be something that is direct.

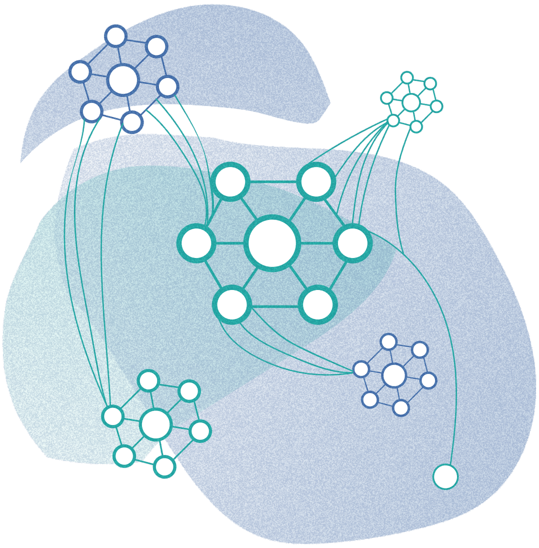
Connected Data is at the core of so many different modern use cases. However it doesn't mean that people need connected data to solve their problem, but rather they want to project connected data into a format that works for them. A good example would be machine learning. The majority of machine learning platforms would like flat, tabular data as the input. The good news is that because the Graph is a Super Structure, it can always "Downcast" to something that is off lesser sophistication, like flat tables.

The trick to become a connected Enteprise and to benefit from connected data is that the Graph should be the master representation of your data i.e. if you can model your data as a graph, then you have the highest level of fidelity of your data that is possible. But what value does connected data bring?

Connected data helps in so much of the overall data pipeline of your business. It helps in cleaning data, deduplicating and merging data, surfacing data with flexible modelling and to describe how data is connected to each other with context. At the heart of CluedIn is a Graph.

So at what point should your company invest in Graph technology?

Well, first we should start by saying that having the highest fidelity of your data is not an easy or quick task to achieve. However once you have achieved it, you can consider that the Graph layer of your data story is your master data source. When we are working with companies, we will typically see that the current status quo includes a Data Warehouse, Business Intelligence tools and a rather manual ETL tool that has been hand rolled or using frameworks like BizTalk, Informatica etc. This is not the high level foundation that is needed to fuel the future of new technology simply because we will only typically get a single document view or a relational / aggregated view of this data. Similar to imagery, you can really only downscale - if you tried to upscale, you will lose some of the resolution and your image would be blurry. The same thing occurs in data where if you tried to upscale from a document or relational view of your data - it would simply get a fuzzier view of that same data.



So really, the answer is that we should have actually started with the Graph in the first place. Unfortunately, this technology did really not exists for quite some time and has only started to mature in the market today. Due to this, most companies will need to take the approach of building a graph out of lesser data structures and because of this, there will be work involved in the upscaling process to make sure we can reconstruct what the data should have been in the first place.

So to answer the question posed in this white paper, “why is no-one doing it” - the Graph database has now been popular for around the last few years. However it has always struck me as a technology that many are still scared to adopt. With talking to developers, we find that many have played with the software but never taken it past that. The interesting part of this conundrum is that it is hard to get a reason why out of these same developers.

It is quite easy to say this in hindsight, but there is no need to worry. The part that is typically complex is that it is very easy to build graphs manually, but you tend to go down a rabbit hole when you are wanting to form a graph automatically out of many different sources. This rabbit hole will lead you to the realisation that before you can do this, you need to clean, normalise, enrich, standardise and many more steps before building an automatic graph out of your data would be a reality. Even after the graph is created, there will always be manual intervention and connection stewarding to take care of strongly typing relationships in data instead of having generic relationships like "Related To".

I strongly believe (and can see the success from our customers) that the companies that adopt the Graph will win in the end.

It is the literal equivalent of searching with AltaVista versus Google - those that backed Google won in the end. One of the wins I would like to discuss in detail is one that plagues all companies. Duplicated data. We work with a lot of enterprise companies and we are often surprised at the simplicity that is applied to being able to identify the same records. There are some systems like file comparison tools that are more sophisticated and will use a plethora of different hashing techniques as to determine if in fact two documents are the same. The same can not be said for determining if two people or companies are the same. Most will resort to simple string comparisons and string distance functions.

Don't get me wrong, these are great and should be included in your analysis, but are in no way a thorough solution. Enter the graph. Imagine the same two pieces of data, but now they are connected in a network or Graph. You can see how the two people are related both directly and indirectly. You might realise that at a metadata level, these records don't show any obvious signs that they are the same but when running the analysis through the Graph it may be much more obvious that they talk to the same people, work on the same documents and this is the "clue" for you to be more confident with the deduplication process. This is one of many techniques where the Graph does a superior job than any other sub structure.

However as mentioned before, the Graph view of data is not necessarily prepared to easily slot into upstream consumers like Power BI, Tableau, Azure ML and more. This data needs to be “flattened” and prepared to be ingested by those systems. This does require knowledge in the areas of Graphs to know how to write queries and perform transformations to project the data as expected, however these languages are becoming ubiquitous and much simpler than traditional SQL ways of querying data.

So we are now hearing more and more the idea of the “connected enterprise”.

There are many ways to interpret this, but at the heart of it is “connected data”. More and more, it is becoming insignificant to know where your data came from (apart from audit and lineage reasons) but rather what is the standardised realisation of this data e.g. It is not important that we got the phone number of a person from Dynamics, Salesforce or Hubspot but rather that we have the phone number and it is accurate. If we received different phone numbers from all 3 systems, it is not important that you know which system said which phone number but rather that you now have 3 potential matches of the phone number of a person.

As soon as we do worry about this, then implementation details leak into the situation and suddenly (and this happens all the time) we go down the rabbit hole of needing to understand the responsibility of all the different source systems. Part of the connected enterprise is not simply about having data in a Graph format, it is about making data available to all parts of the enterprise.

Enter the other interpretation of the “connected enterprise” which is a business that has free flowing data available and easy to access from any department, team or individual. Put bluntly, there are no businesses that have this today, but they are on the rise. What we hope to be able to do is to warn as many companies as possible that free flowing data by itself is actually worse for the company if it is not healthy free flowing data that is clean, accurate, connected, governed, tracked and more.

The way we like to think of it is that to fuel the "connected enterprise" we will then need to arm our end users with a list of tools that are simply "empty shells".

These tools are the brush, they allow the user to paint. The data is the paint. The trick will be that anyone can pick up a brush and paint and I say this in jest, but that doesn't mean that everyone will produce a good painting. This is one of the qualms and parts of self service business intelligence and more is yet to be proven in the market. But I do think that it is a worthy direction to explore, mainly because free flowing data is of no use, if we don't have artists at the end to paint the masterpieces. The friction typically comes today from the fact that while painting these masterpieces, the artist suddenly realises that they need a new colour that they do not have - the analogy being that the data is not prepared in the way that allows them to easily ask for something different. Nothing a simple catalog cannot help with to know what other data is available. The complexity today is that this other data is not connected to the records that you are currently working with and hence a lot of effort needs to be put in place to blend datasources together for this adhoc discovery. What happens if the discovery yields nothing? What happens if more exploration is needed?

This is why the connected enterprise makes so much sense, in that by the time users are wanting to consume data, it is already as blended and joined as possible. If a user was to want to bring in the "Last Actvity" date of a sales deal but there was no way to join that data to the sources you have, then this is something that can quickly be discovered instead of going down the rabbit hole to realise that there is no way to join the data.

There will always be a time where data doesn’t blend, but the idea of the Graph being the super structure is that you have such a stronger chance of it blending if your data is in a network based topology. Why?

Because systems don’t blend naturally. No system is placing in easy ways for it to blend to Salesforce or Dynamics or SAP, it is not in the nature of those platforms, nor the design. Hence systems don’t blend point to point, they blend through a mesh and it might be the 3rd, 5th, 12th or 34th system that you add that has all the missing pieces to blend the entire mesh of systems into a coherent and cohesive network. We like to refer to this as “eventually connected”.

There are certain systems that seem to act as this glue more than others and they typically are very structured and core pieces of the business like Active Directory or HR systems. The good thing with CluedIn is that this does not mean that you need to start with these systems first and design out from there - in fact there really isn’t a lot of upfront designing that is necessary. Rather, it should be realised and expected that just because you have two systems, it does not mean that they will naturally blend together. The solution is not to start going into different systems and manually adding perfect Id’s to join the two systems.

In summary, we are more than confident that connected data is the future and it is not only us that believe this. Analyst firms believe that Graphs will see yet another influx of popularity in 2019 and I for one welcome it.

cluedin