CluedIn

white paper

Does unstructured and transactional data make sense for MDM and CluedIn?

2021 | cluedin.com

Does unstructured and transactional data make sense for MDM and CluedIn?

More than ever, the answer to this question is a resounding yes. The utopian of data is the ability for an engine to understand and correlate data in all types of formats, sizes, structures and more. As data from unstructured sources such as files and mail is processed, essentially the risk factor or fuzzy factor grows in producing a result that is correct. Modern techniques for analysing unstructured data are taking a much more heuristics, fuzzy, natural language approach to being able to understand the structure of unstructured content. Techniques like Named Entity Recognition have given us the ability to identify objects within text. From here they can be used to link to known objects like contacts, companies, dates and more.

If we take a look at files, they are typically made up of metadata and content. In the majority of cases, this metadata and content can be extracted into a raw format that can be processed and understood by machines. In the case of textual content like Text files, Word Documents and PDF documents, libraries can be utilised to extract the content and placed into processing pipelines to handle the understanding of the actual content. CluedIn itself supports 100's of different file formats out of the box including PDF, PPTX, XSLX, DOCX and more. This content is then indexed and structured so that it can be searched from within the user interface, but also so it can be processed in the CluedIn engine to try and extract more meaning from the content.

Just as a simple example, at CluedIn our own CluedIn account contains everything from Social Media content, Mail, Calendar Events, PPTX's, Excel Sheets, SharePoint lists and more. This is because we really want to have a full unified view of all the data across our business. However it can be said that the analytical and aggregated data that we have on our website, social media channels and financial systems is not processed through CluedIn at all. This is partly because we do not currently want to use it - but also because the data can easily be stored in a cheap manner and joined on the way out of CluedIn.

At CluedIn we use a SQL Datawarehouse with Power BI fronting it and the SQL Datawarehouse contains a direct feed of analytical data from the social media feeds and website and the other master data from our CRM system. Service Desk and other operational systems is processed through CluedIn and then joined within the Datawarehouse. This allows us to have the best of both worlds where we can run CluedIn at a good operational cost and we can lookup any single record within our business from a single place.

The value we have achieved from putting unstructured data through CluedIn includes:

- We have truly been able to build a single view of our customers where every mail, excel sheet, document, social media comment, support desk ticket is connected in one single place.
- We can search and find any document within CluedIn no matter where it originated from.
- We can fulfil our GDPR, CCPA and other regulatory needs.
- We can search within CluedIn for documents that already exist instead of having to build new documents and later on find out these types of documents already existed.

As for transactional data, this is always a contentious arguament when it comes to MDM. Some would say that transactional data does not have any place in the MDM world. Our response here at CluedIn is that it really depends on the transactional data itself. If, for example, you have the purchases of all of your customers in your ERP system, it could definitely be argued that having CluedIn ingest, process, clean, correlate and normalise this data could be highly valuable. However, if the transactions were also needed in realtime to be able to process in other systems, you would argue that potentially there are 2 streams of data processing here i.e. one to handle the realtime need and the other to handle the more eventual or delayed use of that data e.g. reporting or system of record.

It must also be noted that CluedIn stores its data in a normalised blob store. This means that if you were to simple want to globally change one value in all records, this is a process that requires the engine to reprocess every single record instead of processing a denormalised reference to a value. This is also something to think about when you are storing large numbers of transactional data in CluedIn i.e. what could be a 1 second mutation in a SQL store, could be a multiple minute or hour operation in CluedIn. At the same time, it is worth realising that comparing a SQL database with what CluedIn is doing is like comparing apples and oranges i.e. they are both solving differents parts of the overall challenge. In this particular case, if you see this type of operation being a common occurence, it may actually be better to store these transaction in another store and simply store the edge to these transactions in CluedIn.

With the evolution of big data, it is important more than ever to analyse what type of data truly needs the attention of CluedIn and its processing engine. At time of writing this white paper, CluedIn is not the type of engine for processing Terabytes of data in any realistic or cost effective time frame. This is not due to technical limitations but rather due to the part of the data chain that CluedIn is addressing. With today's technology it is not hard to fathom and expect that joining two tables together in SQL should be able to be done within milliseconds. If you also asked that same engine to automatically clean, run data quality metrics, fuzzy merge and potentially hundreds of other operations then you would either expect it to take a lot longer or support distributing this query over many different computers to bring it back at a shorter time frame - but at the same time that distrubted nature costs more money to run. Due to this, do feel free to reach out to CluedIn or their partner network to understand if your data would benefit from being processed through CluedIn to prepare it for use throughout your different use cases.