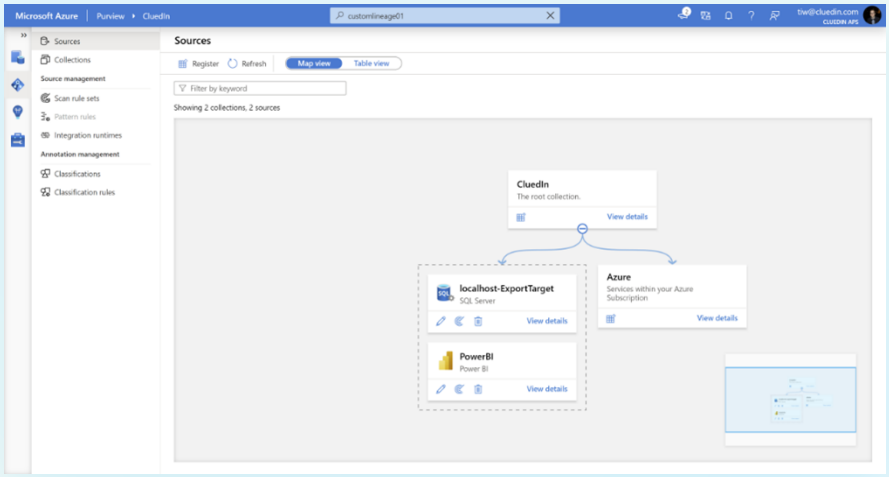
CluedIn

white paper

The Marriage of Microsoft Purview and CluedIn

2022 | cluedin.com

The marriage of Microsoft Purview and CluedIn



With Microsoft Purview now generally available, what better time to talk about CluedIn's native integration with Microsoft's much anticipated unified data governance solution. We have had countless companies ask us how we play with Microsoft Purview and we can say with ultimate confidence, the answer is **"natively"**. We thought that the best way to talk through the integration is to paint a picture of the data challenges that the marriage of CluedIn and Purview solves for enterprise customers.

Imagine a situation where you have **ten systems**. Two CRM systems, three ERP, one Data Lake, two HR systems, Support Desk and Office 365. You have been asked to bring this data together and start generating insights in Azure Synapse and Power BI. This is a common ambition for many companies, but is still quite hard to achieve today. We would like to step you through how this is possible. For this ambition to be met, there are many components in Azure that need to be stitched together. The end goal of this scenario is to generate insights from the data, but to set up a flow that guarantees that insights will continue to be generated as the project evolves.

Let's start by introducing Microsoft Purview. Touted as a unified data governance solution that helps you manage and govern on-premises, multi-cloud, and software-as-a-service (SaaS) data, it addresses a primary objective for many organisations today. It is great to see that Microsoft is bringing this type of solution to the Azure space. Most recently, Microsoft also unveiled the evolution of the Purview portfolio to include important data governance and protection features too.

In fact, it is the era of Machine Learning and Business Intelligence that has highlighted the need for a proper data governance foundation. On top of this, Microsoft Purview is tasked with creating a holistic, up-to-date map of your data landscape with automated data discovery, sensitive data classification, and end-to-end data lineage - enabling data consumers to find valuable, trustworthy data.

Here's how Purview can help answer the following questions:

What data assets do you have?

We should know that there are company tables in CRM #1, and account tables in CRM #2. We should know that we have ten systems, what they are and how to connect to them.

Where is it?

Purview helps us to find the location of the data and register custom metadata in Purview in order to add more details if necessary. It must be said that Purview gives us a framework and shell for interacting with it from its REST API, which means it can eventually answer other questions like:

- What happened to the data from source to target?
- Where is our data being used?

However, we need to do a little extra work to enable this.

Enter **CluedIn**. CluedIn is an **Azure-native Master Data Management (MDM)** system like no other. Throw your preconceptions of MDM away,

CluedIn is a cloud-native, modern solution that unifies data from across the business and prepares that data to generate insights. There are many ways in which CluedIn defies the traditional view of what MDM is. Perhaps the most illuminating is Gartner's statistic that **85% of traditional MDM initiatives fail**. There is clearly something wrong here and a line has to be drawn to separate traditional MDM approaches versus modern. By contrast, CluedIn's failure rate is **12%**. You won't find many vendors talking about failure rates, only successes, but this is just one more reason why CluedIn is different. Transparency.

> With a joint CluedIn and Purview proposition we can finally start to answer different questions that matter to the business. We can also move the needle closer to the point where the data is ready for insight. Adding CluedIn into the mix allows you to answer:
>
> - Who is responsible for data as it moves throughout the business?
> - What happened to the data along the way?

Now that we have learnt a little bit about both the players, let's discuss, at a high level, the main value that comes from the synergy between CluedIn and Purview.

What are the main outcomes delivered by the CluedIn & Purview partnership?

- Purview delivers the birds-eye view, CluedIn allows you to zoom in on the details.
- Purview brings cross-platform visibility into data movement. To recognize this, systems like CluedIn have to remember to register its information back into Purview in order to achieve true end-to-end lineage.
- Purview provides scanning at source, CluedIn provides scanning at record level (but does ask you to move the data in question).
- Purview is your data catalogue for assets across your business, CluedIn takes these assets and turns them into ready-for-insight data.

The first step to data insight, is identifying the data you have which is actionable. You don't want to move and pay processing costs before confirming that the data can be used in the way you want it to be.

Unfortunately, this is a double edged sword - i.e. you often don't know what you don't know. It is also worth mentioning that just like the Data Lake sometimes has a bad reputation as a Data Swamp, so the Data Catalogue is also at risk of becoming something similar if not managed correctly.

There are ways that Purview helps with this. For example, Purview can support Synonyms, so if your tables, files and assets are not all meticulously called 'Customers', you can tell Purview to also look for files and tables called Accounts, Companies and others.

Now that you have registered what data assets you have, there is still a long journey ahead before this data is ready for insight.
CluedIn essentially sits on top of Purview and takes the data to its next step of maturity – think of it as turning scattered data into consolidated data.
CluedIn evolves the maturity of the data by transitioning from thinking about data at the dataset level, into thinking about data at a record level. We think this is necessary because often you will find that your "customers" are scattered across multiple files, tables and systems instead of being in one clean file. Our job at CluedIn is to take the assets and turn them into data that is ready for insight.

At CluedIn, we believe there are two audiences for a Data Catalogue and therefore two types. The first, is someone who wants to know what raw datasets they have across their business. Purview is great for this. The next is someone closer to the business that is not so interested in knowing that customers sit across 4 SQL tables and what the Primary and Foreign Keys are, but are much more interested in the customers themselves, where the column names have been standardized and the data has been normalized.
So now with CluedIn and Purview, instead of four customer files/tables in different formats and structures, we now have 35,342 customers. The job of CluedIn is to align these customers with normalized values and more. That means 35,342 customers that have been de-duplicated, cleaned, enriched and standardized - instead of four tables with customers in it that need that attention every time they are used in the future.

It is important to establish that Purview and CluedIn are NOT the last target for data. This means that ALL tools need to register their movement in Purview, because there is no way for Purview, CluedIn, Azure Databricks and other systems to track data past one hop.

Once a system like CluedIn or Azure Databricks has pushed data to another system, there is no easy way to track what happens after the data has been "let out of the bag". CluedIn not only tracks data coming directly to it (and not just through Purview first), but it also tracks data that goes out to other targets. We write this lineage directly back into Purview as we believe it is the central place that it should be registered. This is how we achieve end to end lineage. It does mean that other systems like Azure Databricks that are handling, moving or storing data, need to support writing data processing and movement back to Purview. Or that Purview reaches out and gets the data from those tools itself.

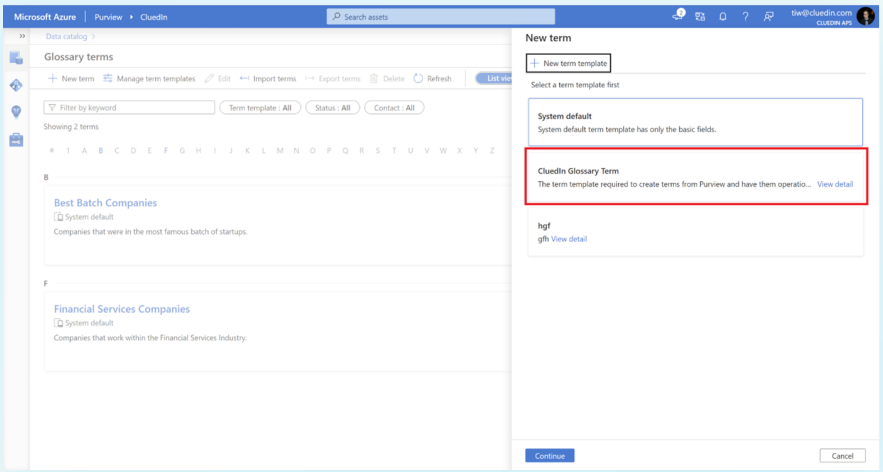
Now that we have Purview serving the data to CluedIn, CluedIn can now consolidate the data and make it ready for us by downstream systems such as Azure Synapse. Let's now talk about how we can slice and dice the data to match the use cases that business want to drive. To enable this, let's take a look at Purview and CluedIn Glossaries.

The Glossary in Purview is about describing assets, the Glossary in CluedIn is about describing data after it has been integrated, standardized and to a record level.

- With Purview, you can answer the question **"Where is my customer data?"**
- In CluedIn, you can now answer the question **"Who are my customers?"**

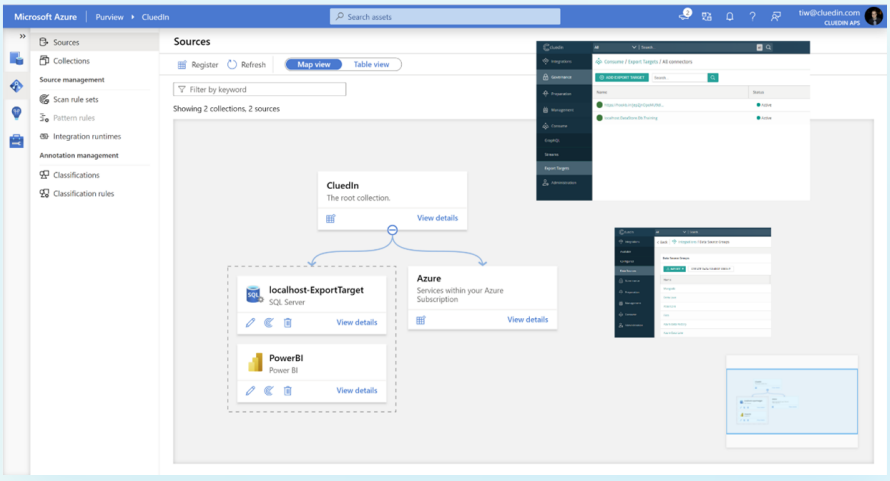
The bottom line is, you need both. You can't have one without the other. CluedIn has worked meticulously to bridge the best of Purview with the best of CluedIn to offer a seamless and fluid experience. Here are some examples of the integration in the flesh.

> **The Microsoft Purview Glossary is available directly in CluedIn and vice-versa. Allowing you to easily transition from an asset level Glossary to a record level Glossary.**

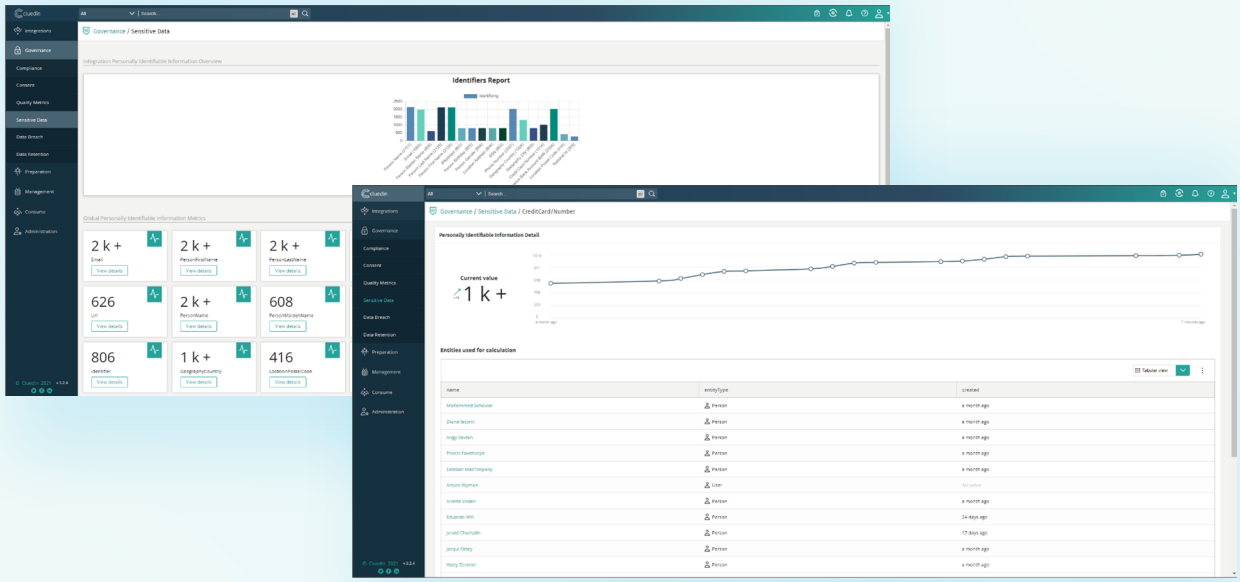


CluedIn can ingest assets that have been registered in Purview.

When registering assets in Purview, you can assign Key Vaults to the resources. Given the right access, CluedIn can read directly from the Purview Registered Resource and the Key Vault to self-authenticate with the source, requiring one less step in the chain of getting data into your Master Data Management (MDM) solution, cleaned, and out in the ether generating insights.

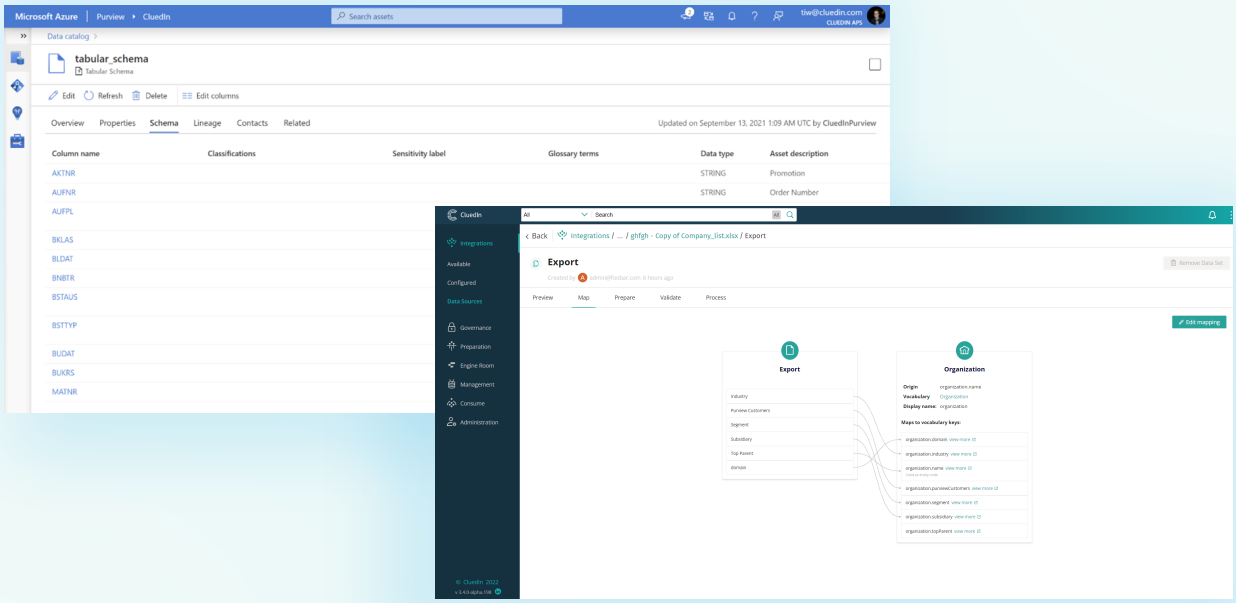


CluedIn scans personal information from Microsoft Purview, and can pinpoint at the record level where that personal data is. It also adds supports for personal information scanning in unstructured and semi-structed data, not just structured. So CluedIn can scan files, mail, PDFs, and more.



CluedIn will use the schema set in Purview to automatically map data sets into CluedIn.

At CluedIn, we have developed a zero-modelling approach to MDM that is revolutionizing the space. In saying this, we can still leverage metadata in schemas to hint to CluedIn what data types and constraints should be used, without enforcing their entry into CluedIn. Rather, we want to flag that there are issues and assign fixes to data stewards to rectify them. Other, traditional MDM platforms will simply reject these records, not even giving Data Stewards the chance to fix them.



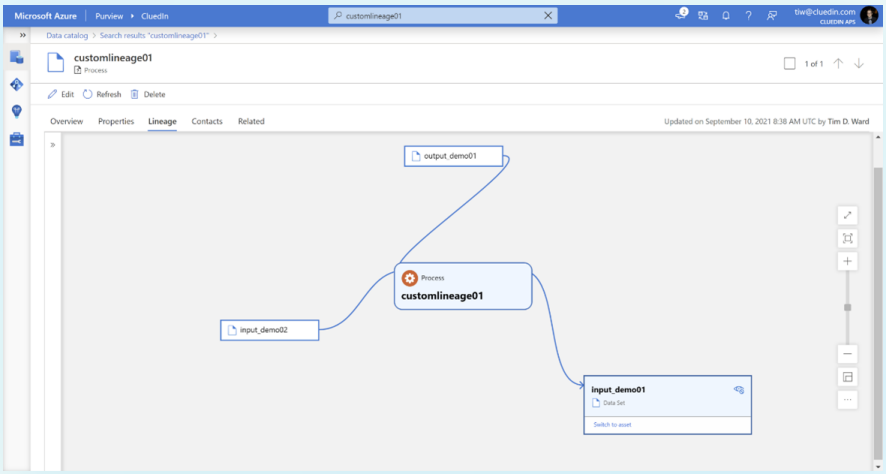
CluedIn extends Purview Lineage with detailed processing logs.

Most companies struggle to track in detail what happens to data as it makes the journey from raw to insightful. CluedIn's job is to explain itself as the transformation take place. This includes, but is not limited to:

- Why did a record merge with another record?
- Why did we choose to use the City from one record over another?
- Where did we get the associated Industry for the company from?
- What business rules were triggered on the data?
- Which IDs did it use to merge with other records?

CluedIn can initiate Purview Scans before a new data ingestion is scheduled.

In some cases, sensitive data cannot be moved between different systems. In these instances, CluedIn can initiate Purview Scans before a new data ingestion is scheduled, adding an extra layer of risk mitigation.



Summary

The combination of Purview and CluedIn is not about achieving 100% data quality across the board before generating insights. It is about putting a system and flow in place to improve data quality over time in a tracked and transparent manner. With this stack you can now answer:

- What data do we have?
- Where is the data?
- What is the quality of the data?
- Who owns the data?
- Who is responsible for each step of the data journey from start to finish?
- What happened to the data as it transitioned from raw to insightful?

Now that CluedIn and Purview have both done their jobs, CluedIn is responsible for making this data available to Synapse (in this case) so that the Data Warehousing team can refine the data in order to generate insights.

Although slightly off topic, it is also worth mentioning some of the other tools in Azure and how they fit into the picture. For example:

> **1: Why can't I just plug Azure Databricks or Azure Synapse into Purview and give it to my Data Engineering team to solve this challenge?**

Of course, this is an option. But there is a good reason why the concept of Master Data even exists. Certain parts of the data management process require involvement from data engineering and IT, but in other parts of the process the involvement of IT would result in an unsustainable workload. This is because some parts of the data processing journey simply do not scale and do not work well if attempted from an IT perspective.

Here are some very practical examples of things that on the surface can be solved with data engineering, but in practice, do not scale.

- We have data on companies coming from multiple places, however the cities are all deformalized. Sometimes people write SYD, sometimes Sydney, sometimes Cydney! On the surface, this seems like a simple “if” statement, but then you have to multiply it by all the possible permutations and all the cities. Let's just say that IT can solve this. The problem is that whenever there is a new permutation, then IT has to get involved again.

- When bringing data together from across multiple sources, it isn’t always obvious how records should be merged. There are great fuzzy merging libraries in Python, but once again, this will need constant attention from IT as anomalies arise.

- When bringing data together from across multiple sources, you will find that datasets 1 and 3 can talk to each other, but 1 and 2 can't. For datasets 1 and 2 to talk, they have to jump through an ID in dataset 3 and 4. Now add one more dataset and you have 5 (!) possible links to investigate to see if records can be triangulated. This is NOT a scalable approach to integration. We can say this with full confidence, as evidenced by one CluedIn specific customer example with 642 data sources. This would be unachievable in ANY other tool.

- You will most likely be using the same code and logic as new datasets come through, but it is unlikely that you will be re-using it properly. CluedIn can centralize automatic transformations so you don't have to maintain dictionaries of values that should be auto-transformed.

Is there any place for Azure Synapse directly over Purview?

Absolutely - 100%. There are a couple of obvious ones, but let's focus on the main one. Processing data through CluedIn takes time and attention. If you require QUICK, but not necessarily real-time, answers about your data and are happy to accept that the data has holes and issues then it makes complete sense to go with Synapse. However, when it is time to make REAL decisions based on this data, it needs to have been matured through a platform like CluedIn. This means that duplicates have been removed, data is normalized and ownership of the data has been established – amongst other things.

We have to remember that the Spark/Python approach to preparing data is a very different workflow to that of the MDM style. Typically, the Spark/Python approach is much more about taking raw data, and scripting notebooks which allow us to see the code being evaluated without our own eyes. It really is a lovely approach to the problem. Once we have solved the challenge, we essentially save these as automated pipelines. In principal there are many cases where this would work beautifully. The problems arise when things change and new problems bubble up. Then you have a choice. You either maintain these pipelines and involve IT again and again every time a new problem comes up OR you systemize the changes in a system like CluedIn. This is in no way saying that CluedIn replaces Spark/Python. This is saying that there are certain aspects of the data journey that should be solved with Spark/Python and others that should not.

Should CluedIn write back data to Purview?

No, because Purview is not a data store, but rather an application that scans data stores. Hence it could make complete sense for CluedIn to write data back into a folder in the Data Lake called "Cleaned".

This does complicate the architecture as you don't have the lovely left to right flow of data where the data becomes more mature, less flexible and more ready for insight. However it does play nicely in terms of some of the more modern approaches of Lakes, Lake Houses and Data Warehouses.

> **Don't MDM systems only support storing Master-style Domains like Customers, Products?**

Yes and no. Most MDM platforms will tell you "Yes", you only store certain types of data in an MDM system. *We strongly disagree with this mentality*. Why? Because, we are asking the wrong question of our data. Instead of looking at Master Data as a short list of Domains, we should be asking "what data needs the attention of Master Data Management?" If we look at it from this angle, we then start to ask:

"What structured, semi-structured, unstructured, transactional data needs to be integrated, governed, cleansed, tracked, enriched and deduplicated". In many situations, we could easily argue that all types of data need this attention. Then we need to ask, is this the BEST place to treat this data. The answer then could be quite different. But this doesn't negate the fact that MDM data just falls into a couple of common buckets. I can categorically say that this is just plain wrong. It honestly sounds like vendor-speak for "ours can't do that, so you shouldn't do it."

> **My Purview Account is scanning Hive Tables, Data Lakes that literally have TB's or PB's of data. How will CluedIn tackle this scale?**

With pure transparency, CluedIn shouldn't. It doesn't make economic sense. Our aim is to innovate in a way that does handle this, it just won't make economic sense to host PB's of data in CluedIn today. It is kind of like storing data in a hot-seat. In saying this, CluedIn can scale immensely, and most of our customers have many, many millions of records. In fact, we have one particular customer with over 1 billion records in it. But just because it can do this, it doesn't mean it should. In fact, in many cases, what CluedIn often highlights for customers is that they don't even use a large percentage of their data to generate insights anyway.