CluedIn

white paper

Understanding Data Lineage

Pages: 4 | Reading time: 9min

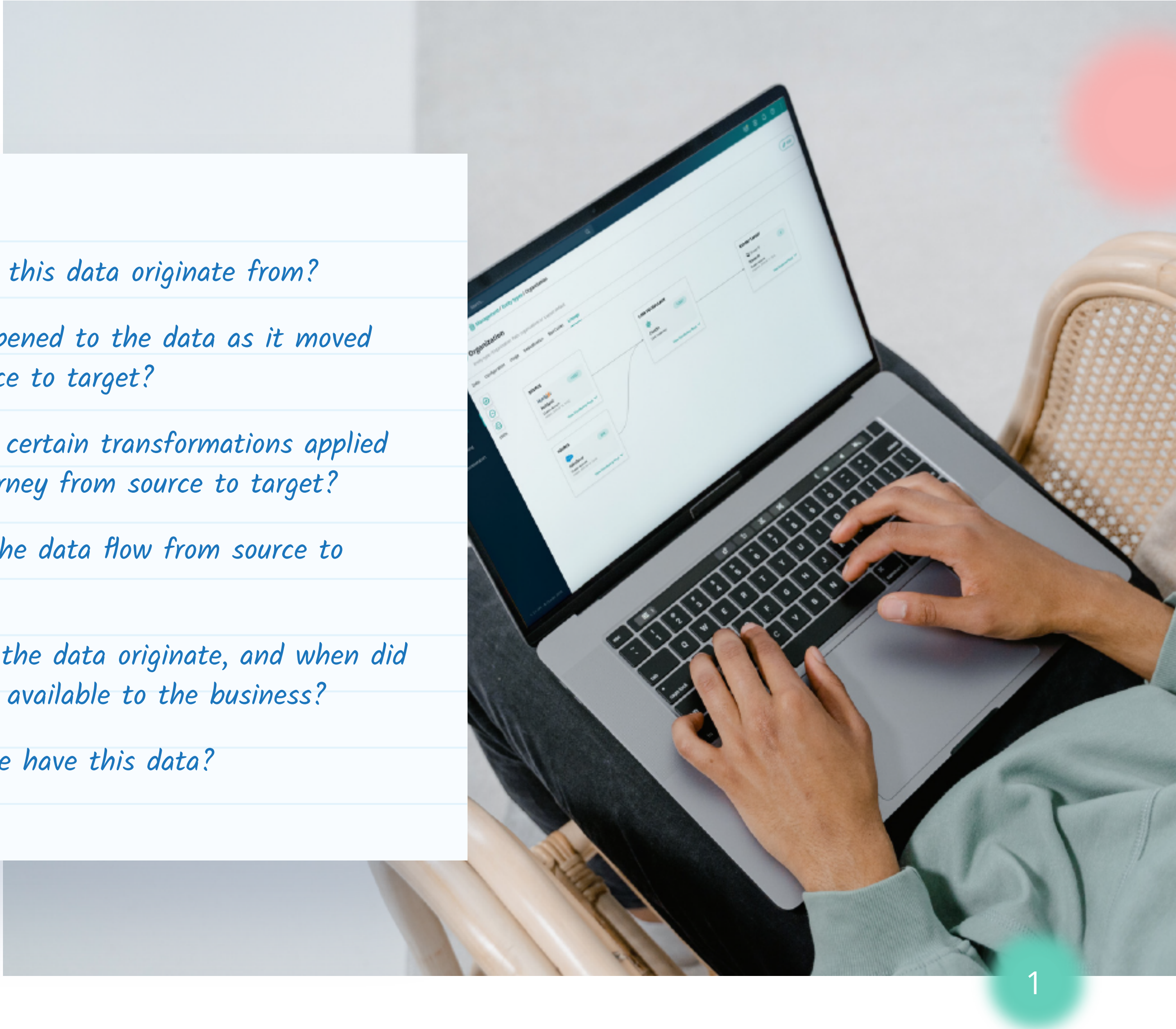
www.cluedin.com

Understanding Data Lineage

What is Data Lineage?

It is widely accepted that data lineage is a crucial component of data governance, data quality, data analytics, data privacy and security. Traditional definitions of data lineage focus on the origin of the data, what happened to it (e.g. Transformations) and where it moves over time. However, in order for data lineage to address the needs of modern organisations – where data is growing in terms of both quantity and proliferation – it needs to offer much more than the basics.

At its core, Data Lineage should address the need to have complete visibility and accountability over the journey data takes as it flows throughout your business. The transparency that Data Lineage provides should allow you to answer questions like:



- *Where did this data originate from?*
- *What happened to the data as it moved from source to target?*
- *Why were certain transformations applied in the journey from source to target?*
- *How did the data flow from source to targets?*
- *When did the data originate, and when did it become available to the business?*
- *Why do we have this data?*

Why is Data Lineage important for my business?

Data Lineage brings trust and traceability. It allows you to assess the credibility of data based on its provenance and the journey it has taken through your organisation. With most modern enterprises relying heavily on data in order to perform analysis and make decisions about everything from product enhancements to go-to-market planning, being able to accurately assess the legitimacy of data is vital to business performance.

More importantly, having this lineage readily available to all, and not buried within some IT system that only a few can actually see, is something that will ease potential bottlenecks in the usability of data.

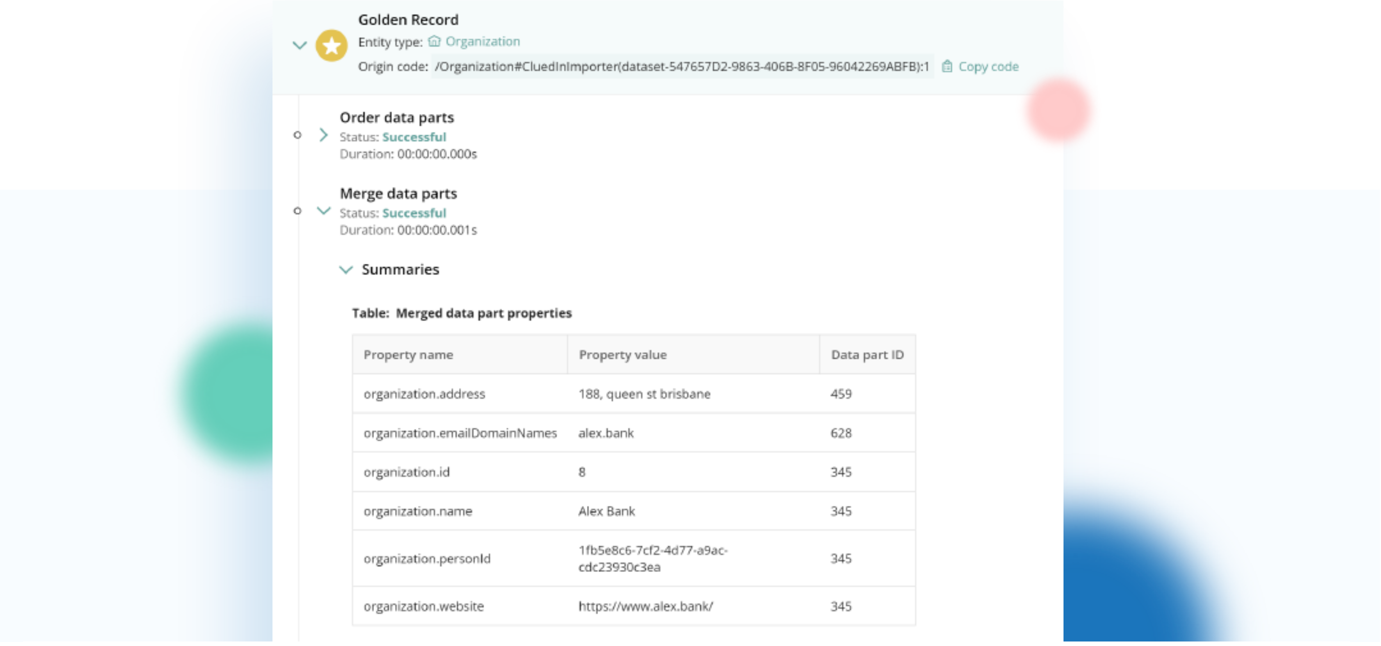
Imagine a situation where you are in a management meeting and are asked to explain the report you are showing the group. Having the ability to answer where you got the data from, what happened to the data on its journey to this report and the transformations that happened to the data, will build confidence and trust in the actions you recommend based on that data. In addition to this, imagine being able to call out who was responsible for this chain, the different parts of the chain and the decisions that were made along the way.

In addition, data lineage can help you to identify and correct the sources of mistakes, and avoid false assumptions which can arise as a result of data bias or undetected data sources. It also allows you to identify and avoid data duplication, which simplifies processes and lowers costs. It is often difficult to answer the question of why something changed. Imagine a situation where the lineage can clearly show that Jess Smith changed the VAT Number for a Vendor from DK123456789 to DK987654321. Although we can see there was a change, without the ability to augment this data with a comment, it will be hard to know why that happened. In most cases, it will be because of an acquisition or a merger, but without comprehensive lineage it will remain an unknown.

One of the other main reasons organisations are choosing to invest more in data lineage is in order to meet regulatory compliance standards and reduce risk. The EU General Data Protection Regulation (EU GDPR) is a perfect example of this as it pertains to the use and retention of personally identifiable information (PII). With effective data lineage, not only can you adhere to regulatory requirements more efficiently, you can also respond to changes more quickly and with less disruption.

How should I approach Data Lineage?

Modern master data management solutions deliver far more than traditional definitions of Data Lineage might lead you to expect. For example, we believe that it is not enough to simply talk about input values and output values. This is why the CluedIn platform offers the “Explain Log”, which is essentially a way of describing the logic of Transformation in an approachable way to a user in order to help them understand why an input value resulted in a particular output value. This requires us to annotate transformations and logic in a way that articulates WHY certain logic trees are generated. It is open, pluggable and easy to use.



Golden Record

Entity type: Organization

Origin code: /Organization#CluedInImporter(dataset-547657D2-9863-406B-8F05-96042269ABFB):1

Order data parts

Status: Successful

Duration: 00:00:00.000s

Merge data parts

Status: Successful

Duration: 00:00:00.001s

Summaries

Table: Merged data part properties

Property name	Property value	Data part ID
organization.address	188, queen st brisbane	459
organization.emailDomainNames	alex.bank	628
organization.id	8	345
organization.name	Alex Bank	345
organization.personId	1fb5e8c6-7cf2-4d77-a9ac-cdc23930c3ea	345
organization.website	https://www.alex.bank/	345

When it comes to answering the question of why something went wrong, you should look for a solution that provides a framework which registers movement, transformation and traceability from anywhere. Imagine a situation where someone asks "Why did that lead not make its way through to our CRM system?". There are literally thousands of things that could have gone wrong. The network could have been down. Someone might have accidentally not selected that record to move to the next step in the processing pipeline. An automation or workflow may have been accidentally been switched off. While it may not be possible to pinpoint exactly where the process broke down because of the number of variables, having a framework that can provide lineage to a high degree of fidelity will certainly narrow those options down and make them easier to fix. The key difference here is that it is just as important to understand why data moved, as it is to understand why it did NOT move.

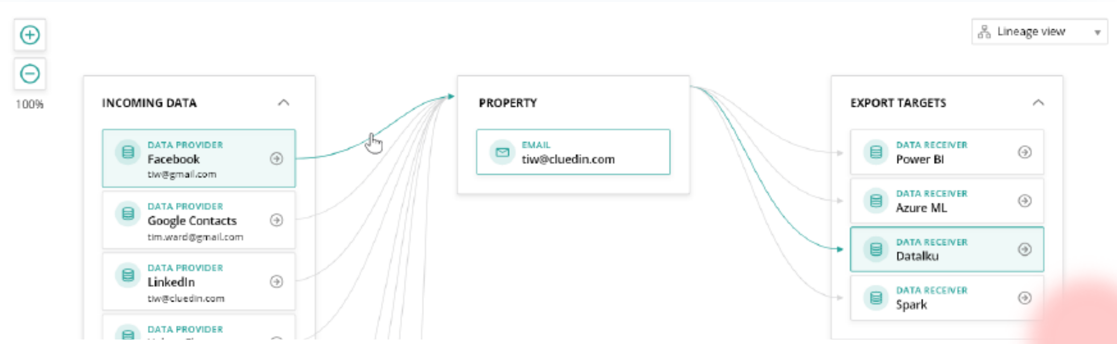
Data often starts its journey long before you became aware of it. Just because you found the data in in a SQL Server Database does not mean that is where the data originated. To trace it back further, you have three possible options:

- *Hope that the Data Lineage has been carried through to the SQL Server on your behalf (highly unlikely).*
- *Use public registries to help data management systems map the original source and previous lineage of the data.*
- *Manually reconstruct it in an effort to clean up technical debt, and add this practice to your future data movement tasks.*

At CluedIn we use the second approach, and are able to pull this information from systems or have it pushed to us. For example, we can reconstruct layers of data lineage pertaining to data that came from Microsoft Dynamics, then moved to HubSpot via Zapier, then made its way to a SQL Server Database.

Data lineage as part of master data management

If the primary goal of a master data management initiative is to create a single source of truth, or Golden Record, for every object within the business, it stands to reason that data lineage should be part of that effort. Mastering your data means having absolute confidence in the provenance and credibility of that data. Without it, it is rendered useless as the basis for analytics, machine learning and strategic decision making. Advanced master data management systems bring together data quality, data governance and data integration capabilities and orchestrate them in an automated way that delivers direct value to business and technology users. Data lineage is a critical piece of this puzzle, and it's worth is only enhanced by being part of a more comprehensive and augmented data management offering.



CluedIn

www.cluedin.com