CluedIn

white paper

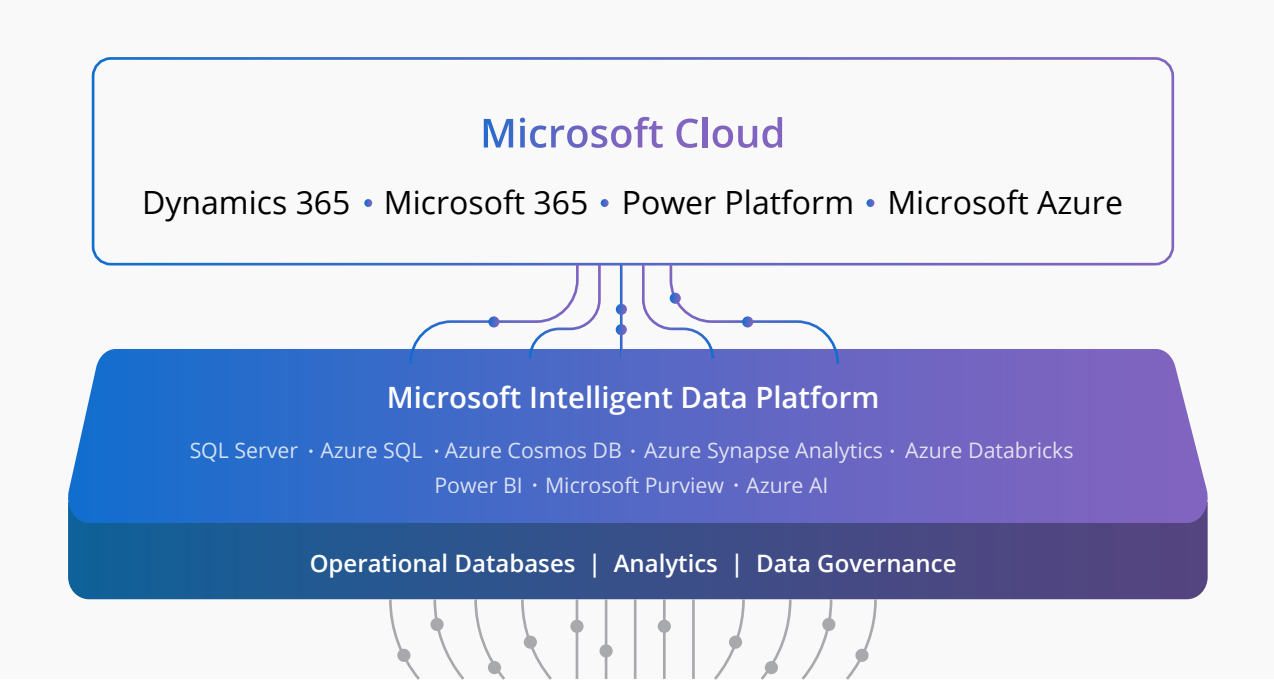
A comprehensive guide to implementing Master Data Management with the Microsoft Intelligent Data Platform

How to build a data foundation in Microsoft Azure to fuel your data-driven ambitions

Pages: 9 | Reading time: 22min

www.cluedin.com

A comprehensive guide to implementing Master Data Management with the Microsoft Intelligent Data Platform



The missing piece of the data puzzle

For companies looking to innovate, minimize costs, and monetize new opportunities, how they drive value from their data estate is a determining factor in whether they succeed or fail. Similar to the speed, scalability, and efficiencies offered by Cloud computing, the ability for companies to do the same with their data operations – although not yet realized for most – will have a similar impact.

The proliferation of Artificial Intelligence (AI), Business Intelligence (BI), Automation, and Cloud has served to highlight that without a proper data foundation in place, these initiatives have great promise but their potential is rarely met unless fueled by the right processes, people, technology and culture.

Fortunately, instead of every organization having to research and build its data estate and operational model from scratch, Microsoft has combined its mass of experience with that of its partners to present a core data management framework that contains all of the essential pillars you need to yield tangible value from data. These pillars are not optional – they form part of a solid foundation that is necessary for every business.

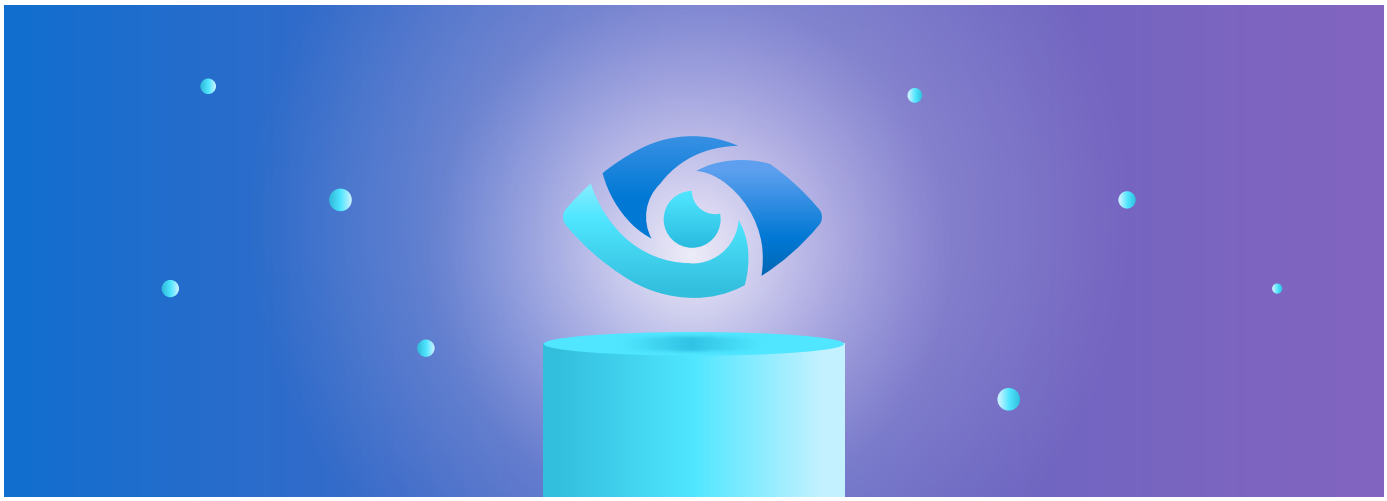
The Microsoft Intelligent Data Platform is a powerful single platform for databases, analytics, and data governance tools. It includes Microsoft's applications and services as well as those of select data management partners – such as CluedIn. These pillars combine to deliver a fully integrated, end-to-end data foundation on Microsoft Azure.

The introduction of the Microsoft Intelligent Partner Ecosystem in October 2022 acknowledged that as beneficial as it was to have its pillars of Data Integration, Data Warehousing, Data Lakes and Visualization tools, Microsoft felt it was necessary to enhance these with Data Governance and Master Data Management systems which could address the requirement to deliver and manage ready-for-insight data consistently and reliably.

Everyone benefits if we play together

Master Data Management (MDM) has many roles to play in the delivery of data that is ready-for-insight, but without the other pillars of the Intelligent Data Platform will likely fail to deliver data that can be used as a strategic asset for your business. Specifically, the role of MDM in the Intelligent Data Platform stack is to provide an operational, high-quality flow of data as it moves through your data estate. This data pipeline should be operated by the business, and it should provide the ongoing, daily maintenance of data quality for you to solve data quality issues as they arise – without having to go back to IT to fix them. IT has its own tools within the Intelligent Data Platform that allow them to solve data challenges using automation and code. For everything else, there is MDM.

It is important that MDM doesn't overlap with other components of the Intelligent Data Platform. It should complement and play its role in the process without duplication. MDM and its relationship to Data Governance is a perfect example of this, as each relies heavily upon the other but has its distinct role to play. Enter Microsoft Purview, CluedIn's companion in the Governance pillar of the Intelligent Data Platform. Microsoft Purview is designed to help you easily create a holistic, up-to-date map of your data landscape with automated data discovery, sensitive data classification, and end-to-end data lineage.



Microsoft Purview answers questions like "What data do we have?", "Where did the data come from?" and "What policies should we enforce on our data?" In many cases, these policies will be a combination of policies you have to abide by and company-specific policies that are unique to your business. For example, today it is not an option for you to choose if you would like to know where PII data resides in your business, but it might be a policy that your Sales team cannot see data generated by your Procurement Team. As the name suggests, Microsoft Purview will give you a bird's eye view of data coming in, data movement, and data going out. However, we need other parts of the Intelligent Data Platform to help generate insights, visualize them and cater to the different types of data we may want to track - such as IoT, operational, or analytics data.

Deciding which task is best handled by which part of the stack

One of the benefits of the Intelligent Data Platform is that it gives you different options when it comes to solving data challenges. The flip side is that you can have too many different ways of solving the problem, and you need to decide which is best.

For example, can I match and merge records within an MDM system? Yes. Can I do the same within Azure Synapse or Azure Databricks? Yes. But why would I do it in one part of the Intelligent Data Platform versus another? It all comes down to Build vs Buy and the day-to-day operations of that solution. MDM is specifically designed to solve a couple of core data issues and solve them well.

These include:

- Merging records from many different systems into a Golden Record with Identifiers, or using Fuzzy Merging to create one record with four entries.

- Enriching and validating records from public data sources or data purchased from a data broker. Systems like Azure Databricks and Azure Synapse rely on all of your data running in memory, and it is considered an anti-pattern to use network lookups in Spark jobs. This then leaves you with building these yourself in Power Automate, Azure Functions, Azure Logic Apps, or custom code.

- Providing tools for ad hoc, non-code-driven data quality fixing. Once again, you could do this in Azure Synapse or Azure Databricks. But data does not stand still and needs to be treated operationally. Therefore data quality fixes performed by Azure Databricks today will not solve the issues you find tomorrow. To stop the vicious cycle of always reverting to IT, MDM provides an operational way to detect and fix these issues "on demand".

It is also important that MDM is not viewed as “just another source”, but rather as part of the operational pipeline. A commonly used analogy is the equivalent of the robots that sit on the conveyor belt and detecting and disposing of the bad berries. It may not be a perfect metaphor, but it demonstrates that some bad berries will slip through, and when we learn more about what makes a bad berry, we can retroactively address that, instead of locking our data up and only letting it out when we believe we have solved every possible issue with it.

You may be wondering why you need an MDM system at all if you have Azure Data Factory and Azure Synapse already. There is a very good reason why Microsoft has made sure to include MDM as a core element of the Governance pillar of the Intelligent Data Platform. Azure Data Factory and Azure Synapse are very good at joining and blending data that is pristine (with no data quality issues and typically coming from the same source systems). The reality though is that your data foundation will need data from across many heterogeneous systems, and though they might join and blend well in each system, they will not do so across multiple systems.

Systems like Azure Data Factory will and should be used during parts of this process, but the blending of records into a unified record is not one of them. Azure Data Factory will and can play a significant role after this data has been run through MDM. Why? Because the merging and linking of records is fundamentally different from joining files and tables from one source system.

For you to have data that is ready for insight, you need to stop thinking about Files and Tables, and rather in Domains. And a particular Domain – such as the Customer – will most likely be constructed from multiple Files and Tables. Without MDM, you will always be working with raw data – i.e. using four customer files and 13 customer tables as the basis for a project. The timelines for these projects can be significantly accelerated if we can serve a comprehensive list of customers directly from the MDM system instead.

The other big challenge with data, hence MDM's inclusion in the Intelligent Data Platform, is that it is often not in a state that is ready to match and link. Real-life data is gnarled, missing values, and unharmonized. Some of these issues can and should be solved by the IT team, but there are also those which should not. Even if the IT team could solve them all, this would mean joining the back of the queue and raising tickets every time. Which would slow down the process of generating insights even further.

Getting started with the Intelligent Data Platform

The Intelligent Data Platform, including CluedIn, is designed to be as flexible as possible and to support your business priorities. CluedIn can be introduced into the Intelligent Data Platform at any point, with minimal impact or rework required. A logical place to start with CluedIn is after Microsoft Purview has been set up and you have your first system scanned, registered, and governed.

Unlike traditional MDM approaches that force you to build your information model upfront, modern systems like CluedIn are led by the data itself, creating a natural data model as the data is ingested. As one of the primary roles of Microsoft Purview is to tell you what data you have, it makes sense to have that in place before embarking on an MDM project.

The next question is whether you should build your Data Warehouse, BI, and Machine Learning (ML) reports before having an MDM system. Modern MDM systems are built to embrace the entropy of a business, and understand that change is constant. This means that systems like CluedIn can be introduced at any point during the project.

While it is perfectly possible to implement CluedIn after the core framework of your Data Lake, Data Warehouse, and BI and ML tooling is established, you will need to be clear that although you can start to serve out data to the business, it will lack data quality control, will most likely contain duplicates and the data will not be harmonized. With this in mind, we know that when we turn on the CluedIn MDM component these issues will be addressed via operational control over the data as it flows to downstream systems.

We would always recommend that you implement your Intelligent Data Platform stack in a way that allows you to deliver value to the business quickly, and increase value over time, instead of holding onto this powerful technology until all components are in place and only then opening the floodgates.

Our recommendation is to start with the Data Lake, the Data Factory Pipelines, the Data Warehouse, Business Intelligence, and Machine Learning, and then layer Data Governance over that with Microsoft Purview and CluedIn. Your Intelligent Data Platform implementation will always be unique to you, and you may also want to have Data Compliance, Data Privacy, and Data Governance set up very early in the process. This is a discussion to be had with the business as often having these pillars shows as much value as a nice shiny report in Power BI.

Implementing CluedIn as part of the Intelligent Data Platform stack

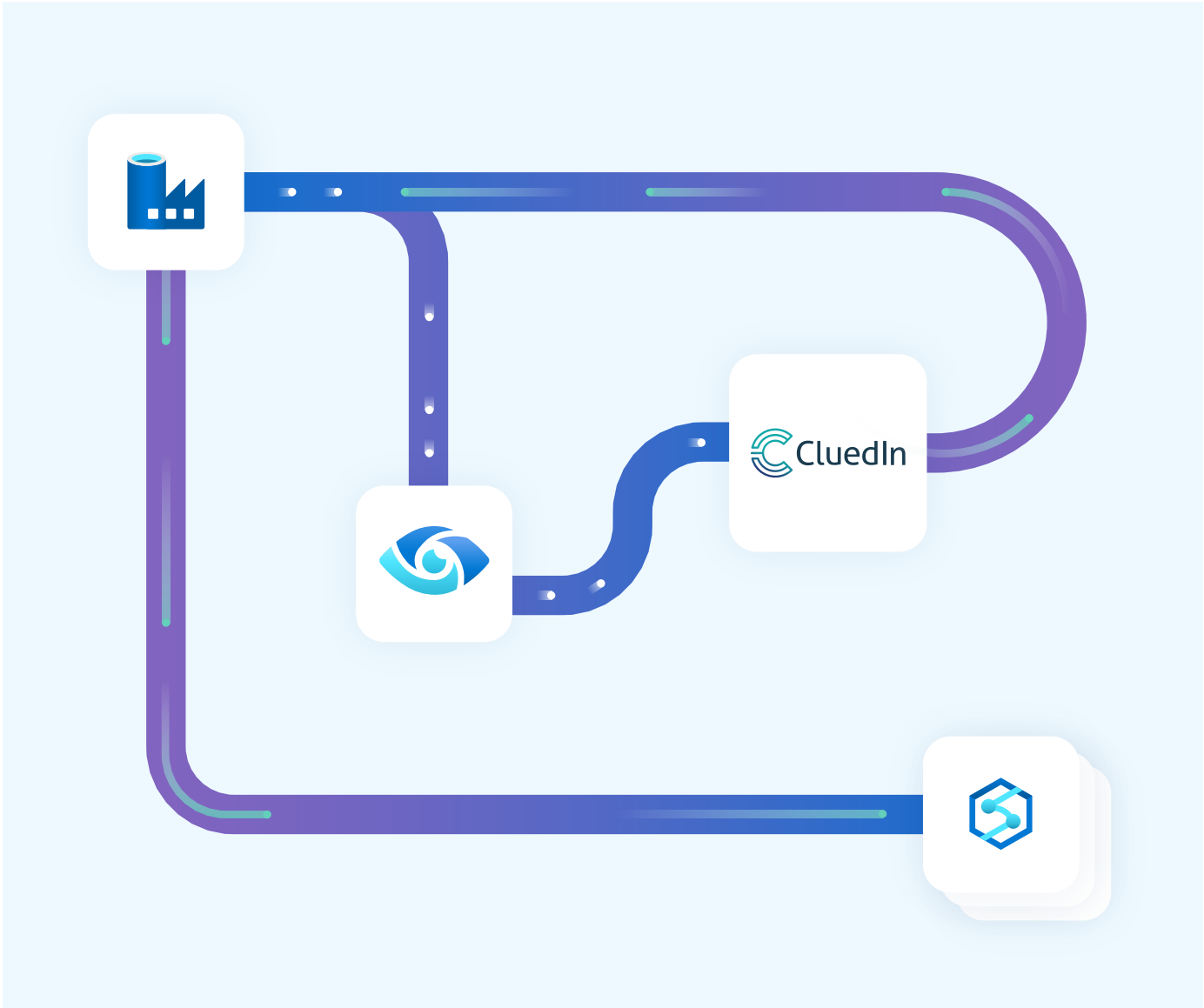
As a starting point for the Intelligent Data Platform, we would recommend using Azure Data Factory to establish a consistent means of moving your operational data into the Data Lake, so that you have one place for your data, in one standard file format.

We then suggest installing your Microsoft Purview instance, in which you can set up regular scans over the Data Lake storage to tell you what data you have in your estate from one single place. Because you are using Azure Data Factory to move the data from the source into the Data Lake, you will have Data Lineage included from the source to the Container Storage. In Purview, you can now start to establish Governance Policies, including defining Business Terms, setting up Classifications for in-place PII and Sensitivity Detection, workflow, Data Sharing Policies, and Data Ownership.

At this point you should install your CluedIn instance, providing CluedIn with credentials for your Azure Data Factory and Microsoft Purview instances. CluedIn can now set up automated data pipelines from the Container Storage to CluedIn, allowing the data to be combined, enriched, mastered, harmonized, and more. Now you can start to enforce your data policies at the granular level.

Microsoft Purview allows you to set up policies at the Metadata level, but you need this at the record level too. For example, you might want to ensure that all phone numbers have an international dial code. This is something that can be established at the Metadata level but can be complemented with a policy at the record level that determines what to do when a country is stored as AUS, UK, USA, etc.

This is also your opportunity to set up the responsibility chain and process around who should fix what. The next step is to set up CluedIn to export this data in a live stream back to your Container Storage (Data Lake) so that Microsoft Purview can also scan and discover it, as well as provide the business with ubiquitous and easy access to data that is now much closer to being ready for insight. With an operational team overseeing the pipeline, the data will mature and quality will improve daily. It is now time to add Azure Synapse, Azure Databricks, Power BI, and Azure ML to the stack for insights and intelligence generation.



Applying your data policies beyond CluedIn

Not all data should flow through your MDM system. The best way to ascertain what data should be handled by an MDM system is not to ask which domain it is, but to consider what data needs operational care – i.e. what data will need some level of human intervention to fix, and can't be 100% automated with code.

You will still need to manage your data policies in one place, which is why CluedIn exposes all of its data policies, as does Microsoft Purview in its REST API, to allow you to easily enforce data policies within your Azure Databricks and Azure Synapse jobs. CluedIn supports this directly within other Azure Services like Azure Functions as well. Enforcement of the policies set out in your Data Governance and Master Data Management pillars need to be extended outside of the Microsoft Purview and CluedIn tools themselves, and this is how it is achieved across your entire data estate.

It isn't possible for any one person to remember all of your data policies and when they should be enforced - e.g. all countries should be stored in an ISO format. What is important is to have a stack that has an early warning system built in and that can fix it once, but utilize it across many use cases.

Why is the MIDP a foundational leap for CluedIn and Microsoft?

As well as bringing together the core pillars of unified data management, the Intelligent Data Platform is comprised of tools and systems that complement each other without duplication. For example:

- CluedIn has powerful and granular calculations of data quality across 17 different metrics and writes them directly back to Microsoft Purview.
- Microsoft Purview writes all its metadata and lineage directly into CluedIn.
- CluedIn utilizes Azure Data Factory to move data from its origin to the MDM platform so that you have data movement lineage within Microsoft Purview.
- CluedIn reciprocates by writing all mapping, merging, and linking lineage back to Microsoft Purview.

All of this means that the Intelligent Data Platform is akin to a hivemind, in that the benefits and value brought by one system or tool cascade across the stack due to the high levels of integration between them.

Fast-tracking the data-driven journey

For many organizations, starting their data-driven journey is a challenge. One of the main advantages of the Intelligent Data Platform is that it accelerates the process by making it far easier to select the technology and different elements of the architecture that work well together. In the past Microsoft and its partners would have made these recommendations based on the experience of working with each other; now this wealth of knowledge and skills are brought together in a more structured and formalized way.

By ourselves we are vulnerable, but together we are strong - this is the ethos behind the Intelligent Data Platform and CluedIn is committed to remaining strongly aligned to the platform and to being the most native MDM solution on Microsoft Azure.

Beyond the Intelligent Data Platform

The Intelligent Data Platform offers the core pillars you need to build a managed, governed, and secure data estate. But there are many other parts of the Microsoft platform that can be leveraged to add extra value in the form of low code tooling, enhanced security, governance, and connectors. CluedIn already integrates with the Microsoft Power Platform and supports the Common Data Model,

which can be used to dramatically accelerate your journey to building low/no code applications, and automated workflows with Power Apps or Power Automate. CluedIn is so much more than a native Azure MDM platform, as it is integrated across the entire Microsoft estate, including Microsoft 365, Microsoft Dynamics, and more. We're ready for your data-driven future to begin, are you?

CluedIn | www.cluedin.com