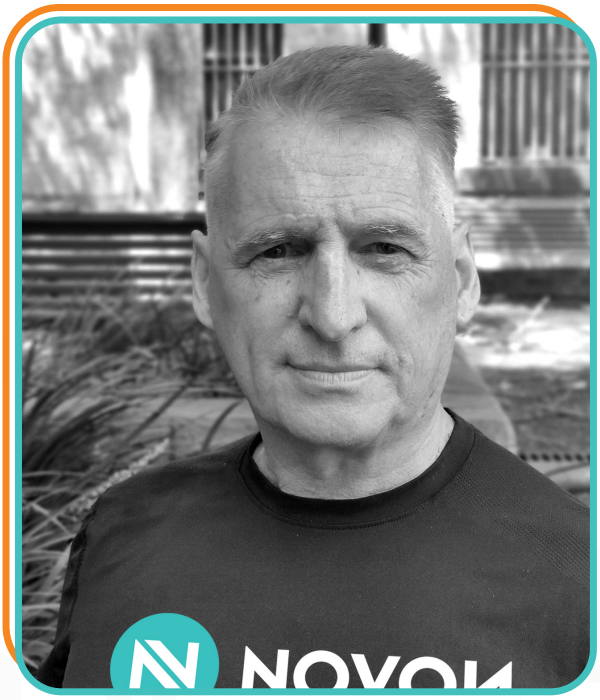
NOVON

Next Gen Data Governance

# AI Automation / Data Catalogue

novon.com.au




# About the author: Christopher Spring

## TECHNICAL DIRECTOR (DATA ADVISORY)

Chris Spring is an accomplished professional with a diverse background in the field of technology, data, logistics, and finance. With a passion for innovation and a strategic mindset, he has made significant contributions to various industries throughout his career. Currently, he serves as a technology executive, leading Novon's data advisory practice, driving modernization and transformation of data assets for a variety of organizations across diverse set of industries.

Chris's has completed advanced studies in engineering, computer science and accounting. His technical background has enabled him to navigate the ever-evolving landscape of technology with ease. He possesses a keen understanding of cutting-edge technologies, such as Artificial Intelligence, Machine Learning, and their applications in solving complex business challenges.

With over 37 years of experience leading strategy and execution of data management concepts, Chris has a passion for Data Governance and is recognized for his contributions on this subject within Australia.

As a thought leader in the tech industry, Chris frequently shares his insights through speaking engagements, industry conferences, and thought-provoking articles. One of his passions lies in Data Governance and enjoys knowledge sharing and actively contributes to the professional growth of others through mentorship and coaching.

Over the years, Chris has gained extensive experience working in leadership roles for both established companies and startups. He has a proven track record of successfully aligning technology strategies with overall business objectives, resulting in enhanced operational efficiency and revenue growth. With his exceptional ability to identify emerging trends and opportunities, he has guided organizations through periods of disruptions and positioned them at the forefront of their respective industries.

In addition to his technical acumen, Chris is known for his exceptional interpersonal skills and collaborative approach. He excels at building and leading high-performing teams, fostering a culture of innovation, and driving cross-functional collaboration. His inclusive leadership style encourages creativity, diversity of thought, and a strong focus on customer satisfaction.

Chris Spring is a visionary leader who combines technical expertise, business acumen, and a commitment to societal betterment. With his forward-thinking approach and relentless pursuit of excellence, he continues to make a profound impact in the technology landscape, inspiring others to embrace innovation and drive positive change.

# Table of Contents

# 1 Introduction

This document's intent is to take you on a journey through the Data Governance arena as of 2024. It will enable you to make informed decisions about the best way to invest in the most appropriate technology for your specific requirements. We have investigated data governance and its expanding requirements over the past 3 years as a result of the digital transformation, and the acceleration this has caused in both, the volume of data and the multiple locations of data due to the expansion of cloud solutions. Once you add the significant increase in hacker activity and subsequent breaches in corporate Australia, data governance is now, well and truly an immediate imperative.

## 1.1 STATUS 2024

Today, Data Governance is considered by many organisations as a difficult to use, rule based, top-down data control mechanism that is difficult to implement and when implemented the end results and benefits are often questionable.

However, when we consider that in Australia in 2022, Optus and Medibank were hacked and millions of customers, related personal information was exposed and ransomed in the market, perhaps there is a strong argument for data governance in 2024 and beyond. What price does a company put on their reputation and the privacy of their customers' personal information?

How do you mitigate this risk but not slow down your core business such that growth and customer satisfaction is compromised, by an overbearing data governance program?

## 1.2 AN EXPLOSION OF DATA

The digital transformations dynamic alone has caused an explosion of new raw data within an organisation. For transforming organisations, it is imperative to understand and disseminate this data so that the effect is positive and rewarding for all sectors of their business. Applicable to all industry verticals, from government to finance to logistics to energy organisations and others.

## 1.3 COLLABORATION, NOT CONTROL.

Data governance shouldn't be something that the human component of data need fear. At its heart, data governance isn't about control. It's about helping data teams work better together. Novon think it's time now look at why data governance is viewed this way, and how we might reframe how data governance is viewed and by doing so save the reputations of data governance and its stewards across your organisation.

At its core, data stewardship, and data governance, was all about collaboration and democratisation. data stewards acted as a bridge between people and process but today data governance is about control, not collaboration.

*Collboration is key.*

## 1.4 RESET THE BRIEF.

So how do we return data governance back to its original core function. Novon believe the answer lies in both the tools chosen to manage the data including the meta data as well as the mind set of those managing the data governance framework. If you look at the progress in nearly all data managed related tools, they have leaped ahead in functionality along with the explosion of data, but not so for data governance. It is still for most vendor tools, a repository of metadata and manual updates to keep it current, which rarely if ever happens after the finish of the initial project.

## 1.5 NEW TOOLS REQUIRED.

Novon believes the answer lies in the tools selected by an organisation to perform this function for them and the mindset of the people involved in providing and managing the governance framework within your organisation.

Into this landscape we introduce the active metadata catalogue tools that use machine learning (ML) to automatically learn and continually update the data about the data, after which it stores this data in an easily accessible data lake, for use by everyone. Gartner has labelled this new tool category, Active Metadata. Throughout the industry this sector has also been called V3.0 data catalogues.

Novon believes this new data catalogue type can return data governance to a collaborative framework. Driving the adoption and use of metadata to improve the timeliness of provisioning data, by adding security features and attributes to your data not possible before and provide additional layers of privacy to your data where required.

## 1.6 WHAT TO FOCUS ON, BUT WHAT COMES FIRST – THE DATA

Novon believes that the first focus in data governance should be about the data and not the people or organisational related components of data governance. It's not that the people and organisation components are not important, but they can be time consuming and often delay the successful completion of a data governance initiative.

Nearly all the data related data governance activities can be completed by implementing a data catalogue, (preferably V3.0 catalogues that includes ML embedded as part of the tool), thereby ensuring that the collection, the curation and profiling of the data need only occur once and thereafter regular monitoring and tuning is only required.

Traditionally, we would have used the DGI data governance framework as our reference. This implies a top-down approach, but as we have discussed in previous sections above, data governance should be collaborative. Rather than following the traditional approach of implementing all 10 components consecutively, please see following a different ordered approach.

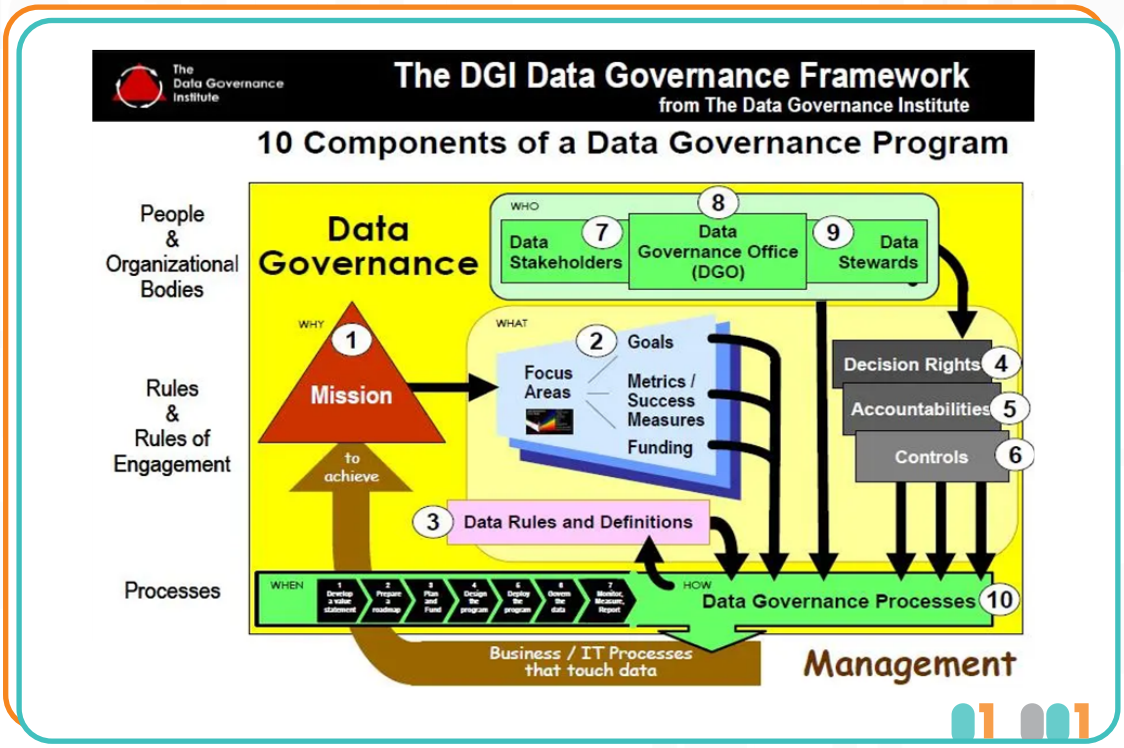



*Figure 1 - Traditional - DGI Data Governance Framework* *courtesy of DGI*

1. Mission and Vision
2. Goals, Governance Metrics and Success Measures and Funding Strategies
3. Data Rules and Definitions
4. Decision Rights
5. Accountabilities
6. Controls

### People and Organisational Bodies

7. Data Stakeholders
8. A Data Governance Office
9. Data Stewards

### Processes

10. Proactive, Reactive, and Ongoing Data Governance Processses

We would adopt a more collaborative approach. This means that five (Nos 1, 3, 4, 6, 7 & 10) of the ten components of the framework are fully implemented first as part of a V3.0 data catalogue implementation. Additionally, number 7, which is the data stakeholder component for data governance is partially completed as part of the data catalogue implementation, however it will need to be extended when completing the other final four components.

The other four components can be resolved after, or if time and budget permit, alongside the implementation of the V3.0 data catalogue, please see the diagram below showing the modernised DGI data governance framework.

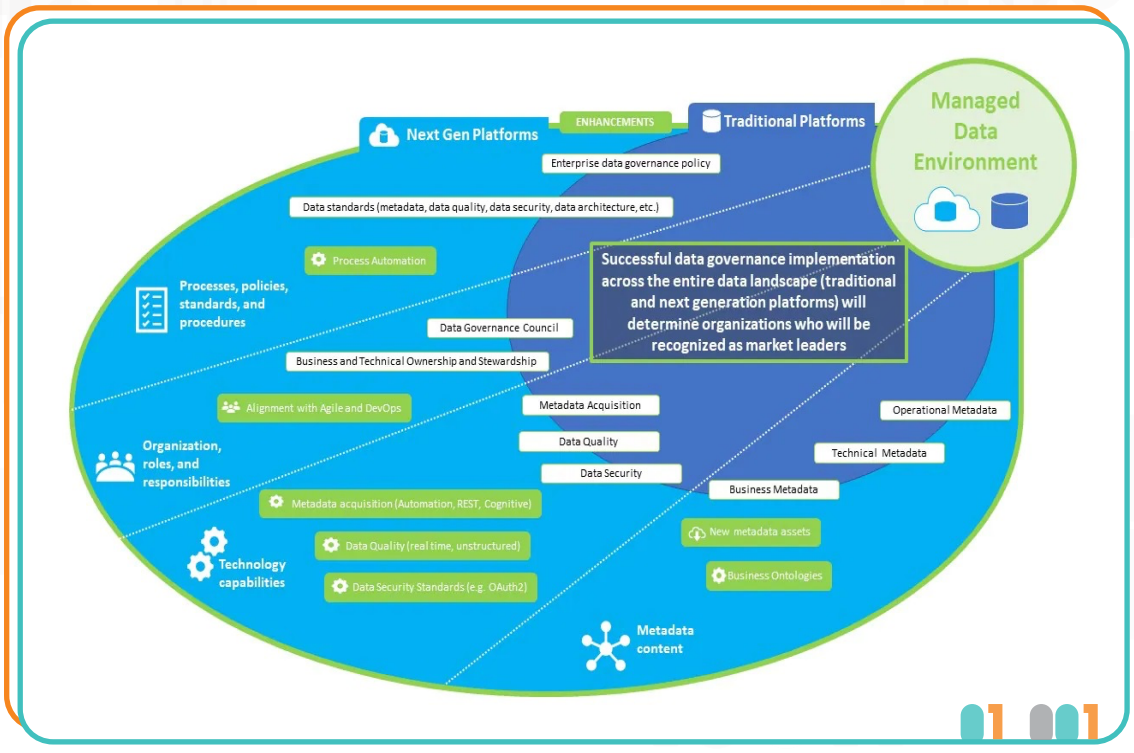



*Figure 2 - Collaborative - Data Governance Framework* *courtesy of Deloitte*

## 1.7 WHAT CAN YOU EXPECT AT THE END OF THE FIRST PROGRAM INITIATIVE?

A program where 5 – 6 of the 10 steps have been implemented for data governance, by completing the data catalogue implementation. A solution that delivers: a sophisticated search engine capability, data cleansing, data deduping, suggested naming conventions, suggested data dictionary, data lineage, tracing of source data applications, and data profiling. The solution attributes and operations can all be defined initially and then continue to operate seamlessly in the background.

## 1.8 WHAT WILL THE FUTURE LOOK LIKE, WHAT ARE THE TRENDS?

Cloud data governance is a must today. As many of a modern organisation's applications exist in the cloud, data governance must be able to handle on-prem and the cloud data requirements simultaneously. Even though AI used in a data governance application is still in its infancy, it is adaptive AI that will revolutionise data governance and cause a dramatic and positive impact on your business in the future.

Data privacy issues were hot news in 2022 and unless we do something different, will continue to be so in 2024. Governments are releasing new legislation to mitigate the issue, without modern data governance programs in place how will you manage this ever-changing and high-risk data landscape. To help in this changing privacy and security landscape, you risk can be managed by a move away from the traditional top-down approach to data governance and move towards a bottom-up approach and data democratisation.

Real-time data pipeline use is on the rise, the V3.0 data catalogues will eventually accommodate this data requirement, but older versions will find it challenging. Real-time capable data catalogues can harness automation that will boost productivity, uncover insights faster, and better manage complex variables. It just requires the right choice of platform.

Older data governance models can tell us what has happened to the data.  However, in 2024 we need to know what is happening (real-time) and not very far into the future we need to know what could happen (predictive), before it occurs, these phenomena will be explored in sections (four and five) of this document.

Novon believes that metadata management will always be the core of data governance.

# 2 Manage the Data explosion, data diversity, data security and data privacy.

## 2.1 TRENDS IN DATA

Let's look at some current trends in digital transformation, data management and data governance that might help us solve this challenge of exponential growth in data and the elevated risks of the data being hacked and exposed. Novon believes in the value of data governance, both to the enterprise and to data management. This value is evidenced in two of the most significant trends to shape this discipline since early 2000, compliance and timeliness of data acquisition.

Firstly, vendors specialising in access management and enterprise security are now concentrating on regulatory compliance, data privacy, and data protection, which has rapidly become a key driver for IT and the business together. Getting this wrong has serious consequences in today's hyper-regulated and escalating cyber risk environments.

The second trend is the real-time application of data governance to encompass a widening array of real-time circumstances, use cases, and market conditions, all of which require data governance to become more adaptable and more responsive than ever to fit these growing requirements in this current market.

Organisations are acknowledging the difficulty of attempting to determine in flight, and or in advance, every possible data governance contingency and prepare for it accordingly. They are looking to vendors to provide new data governance constructs so they can dynamically adjust to new situations as they occur.

### Regulatory and Privacy compliance

More countries are adopting and enforcing data protection and privacy laws that are being passed as legislation in most countries. Companies will need to comply with this evolving regulatory landscape or suffer increasing risks to their business. Data governance will be critical to the integration of privacy regulations and security measures to ensure that data is defined and properly managed for compliance with these new regulations.

### Cloud Adoption

Organisations are moving to cloud environments and need to have data governance solutions that work across multiple environments. Organisations require solutions that will help them transition and migrate to a cloud environment, maintain data governance in a hybrid environment, and fulfill data governance requirements for all data, everywhere.

## 2.2 DATA EXPLOSION

Over the past decade, there's been a significant focus on how to collect more data, primarily due to the digital transformation imperatives and the evolution of the Data Lake. Business leaders, data specialists and data scientists have all wondered how we could collect, store, and present more data. We've presumed that more data equals more helpful information and a better chance to gain knowledge. But this line of thinking has in some cases led us down a deep rabbit hole where we find ourselves today, potentially drowning in data we can't harness.

We've begun to realise that having so much data is not necessarily as valuable as we thought, in many ways, it's becoming more of a liability than an asset. But there are ways to approach our sea of data so that we can leverage its true power. The bottom line is that too much data results in too much noise and compromises the performance, profitability, and security of any enterprise. It can also hide the effective data, making the good data harder to find and leverage.

### So, what are some of the key data statistics for 2024?

- Poor data quality costs the US economy up to $3.1 trillion a year.
- Google processes 8.5 billion searches a day.
- The world will produce slightly over 180 zettabytes (1 zettabyte equals a billion terabytes) of data by 2025.
- 80-90% of the data we generate today is unstructured and most business admit they need to manage this data better.
- According to big data statistics, cyber scams have gone up 400% at the start of the pandemic and the ACSC identifies cyberspace as a future battleground requiring robust protection.

### So where is the data?

While data is growing in volume, the nature and location of data is also changing. Alongside structured data like relational and transactional data in SQL databases, there has been a meteoric rise in unstructured and semi-structured data, and big data, which has altered the data landscape.

IDC predicts that 80% of global data will be unstructured by 2025 because the way we use and consume data, and what we expect of it has changed. Rather than data being stored in fixed, known locations which can be controlled and managed easily, data is, literally, everywhere:

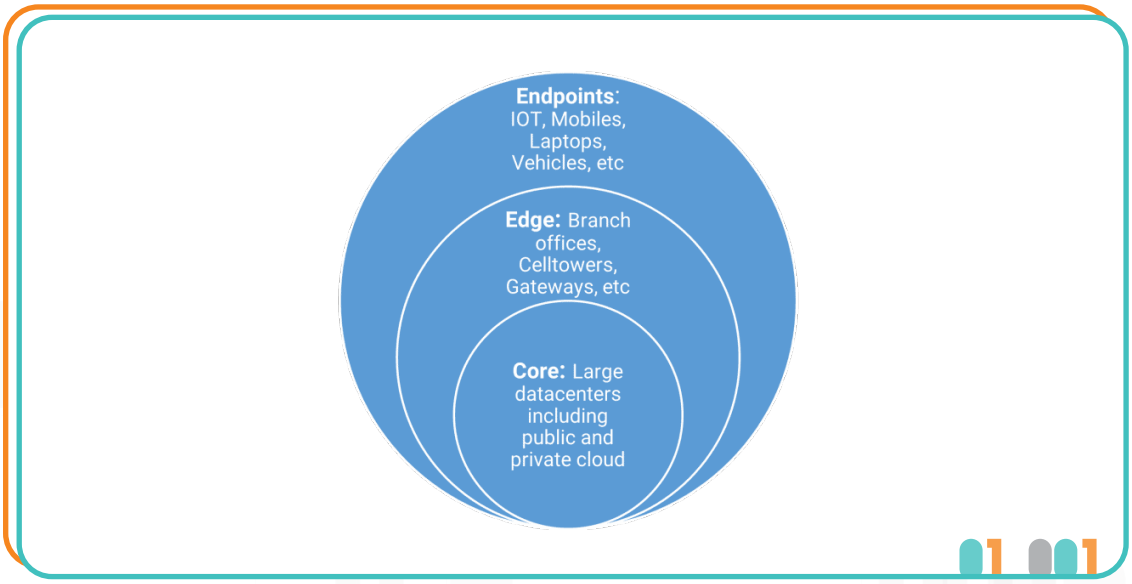



*Figure 3 - Data sources 2022* *courtesy of IDC*

**What to do?**

Many enterprises are now focused on establishing a new technology layer that ingests all their data (ML driven technologies) and, through a set of data processing algorithms (bots), turns it into meaningful information that moves the needle for the business. Novon sees this as the future to handling future volumes of data and believes the humble data catalogue empowered with AI / ML tools will enable this on a scale not previously thought possible.

## 2.3 DATA DIVERSITY

Structured data, unstructured (pdfs, etc.) content, data from IOT devices - Machine code, log and activity data from Spark, Slack, Splunk, and NoSQL sources are only some of the data diversity elements that must be managed in the modern organisation.

A decade ago, there was a belief that tables were the only asset that needed to be managed. But that's completely different now. Nowadays, BI dashboards, code snippets, SQL queries, models, features, and notebook type data are all data assets. In some cases, the generated metadata may provide more meaningful business insights.  The 3.0 generation of metadata management will need to be flexible enough to intelligently store and link all these different types of data assets in one place.

## 2.4 DATA SECURITY AND DATA PRIVACY

There are many data security and privacy laws that organisations must comply with. Novon has listed some of the more common and demanding laws that Australian enterprise organisations need to comply with daily as shown below. Across State boundaries there is also notable variation in local regulations. Obviously, this impacts the governance of your data and the extent to which you extend the attributes of any data entity to support all these privacy and security laws. By no means is this list comprehensive, but are some of the more well known.

- GDPR Europe - The General Data Protection Regulation (2016/679, "GDPR") is a Regulation in EU law on data protection and privacy in the EU and the European Economic Area (EEA) that came into effect in 2018. The GDPR is an important component of EU privacy law and of human rights law, in particular Article 8 (1) of the Charter of Fundamental Rights of the European Union.
- The Australian Privacy Act 1988 (Privacy Act) and amended in 2022 - The Privacy Legislation Amendment
- Basel 3 – Capital requirements for banks worldwide.
- Open banking - Open banking gives Australians the ability to share their banking data with third parties that have been accredited by the ACCC. This allows them to get better-suited banking products and switch products or banks more easily.
- Sarbanes Oxley act of 2022 - is a US federal law that mandates certain practices in financial record keeping and reporting for corporations listed on north American exchanges.
- KYC – Know Your Customer (KYC) is a standard in the investment industry that ensures advisors can verify a client's identity and know their client's investment knowledge and financial profile.
  - Three components of KYC include the customer identification program (CIP), imposed under the USA Patriot Act in 2001, customer due diligence (CDD), and ongoing monitoring or enhanced due diligence (EDD) of a customer's account once it is established.
- SOCI - The regulation of critical infrastructure under the Security of Critical Infrastructure Act 2018 (the SOCI Act) now places obligations on specific entities in the electricity, communications, data storage or processing, financial services and markets, water, health care and medical, higher education and research, food and grocery, transport, space technology, and defence industry.
- There are many other data regulatory standards that organisations need to adhere to in Australia, notably health and state / federal governments as prime examples.

It is understood that the US is preparing a new privacy act after the EU has launched the GDPR in 2018. Some states throughout the US have begun rolling out their own iterations of data privacy regulations. These US regulations are largely focused on the protection of consumers by protecting their rights to decide where their personal data resides, which entities have access to it, and to restrict the sales of personal consumer information. Currently California, Colorado, Connecticut, Utah, and Virginia have active privacy laws, many of which go into effect in 2024.

The information and privacy regulations are indicative that the industry is standing at a technology threshold that started about 8 years ago, albeit quietly and slowly. Driven by increasingly sophisticated user requirements that have emerged over time, and that would not tolerate the existing data governance cumbersome top-down controlled approach, it was clear that new data governance tool sets were needed. A recognition by vendors of the need for the same additional requirements, and by Forrester, Gartner, and others like them recognizing that data governance and metadata management has failed to live up to its promise.

The reason? Data governance and metadata management was passive, and today data management and governance requires active metadata management. In addition, if you have implemented a data governance program previously you will remember the difficulty of gaining agreement and buy-in on data governance, for all the reasons stated in section 1 of this document.

Rather than define a new market, the metadata management market has been redefined and repurposed. This redefinition has caused the data governance and data catalogue vendors to start to extend and improve their offering to provide for these new requirements, but right now in early 2024 there is still quite a way to go to satisfy all the active requirements in the best possible way for the user community.

For Novon, this suggests a two-step approach to achieve the ideal position of best data governance and best data management principles and will most likely span at least 3 years as the vendors continue to add features and function to their current product offerings. This will take some time and if your current vendor doesn't evolve you will need to migrate to a new vendor or risk not meeting the described challenges.

# 3 Where to start, what to do first?

For more than a decade, data catalogue solutions have been the key for bringing awareness and transparency to the data available within an organisation. They help categorise data, assign it to an owner, mark it with a quality score, include protections for security and privacy, and document important metadata. Without the components of a data catalogue, organisations suffer through bottlenecks and poor-quality data making its way to production.

Today, access to a data catalogue and its metadata provides many advantages. Assuming all data that is worthy is captured and documented in detail, finding data becomes much easier. Most of us can relate that a lot of time and resources can be wasted trying to find the right dataset.

Data catalogues can help us determine what data we're missing by documenting what is known. They also make data reuse possible. It's always frustrating when organisations recreate data sets that already exist. In addition, duplicate datasets can inadvertently create integrity issues because it lowers confidence in knowing which dataset is current and relevant.

A data catalogue is the inventory of all the data the organisation has available. It is automatically generated through integration with the data stack tools. A stable, proper data catalogue connects to all elements of the data journey platforms including the:

a. Data Sources
b. Data Lake
c. Data Warehouse
d. Data Models and
e. Data Visualisation tools

Historically data catalogues have allowed us to categorise and secure our data based on its taxonomy, ensuring the right people are getting their hands on it. They have provided an owner and a domain expert for every dataset. This way we know who to reach out to with any questions about the logic, definitions, or quality, they also have a feature that lets us know which datasets are ready to use and which are broken. And currently, they will specify the nitty-gritty of our data such as specific definitions and metadata of our attributes / columns.

Novon believes that metadata management will always persist as the core of data governance.

## 3.1 THINK DIFFERENTLY.

Take on board a new data culture. Business and IT need to collaborate on what they both want and expect from data governance, then the IT folk need to provide a conceptual and logical architecture that uses the 10 steps to complete data governance, but in this new paradigm the steps to be completed are decided for each project, there is no predefined order, rather the order is decided by the data driven imperatives of the organisation at the time of the project.

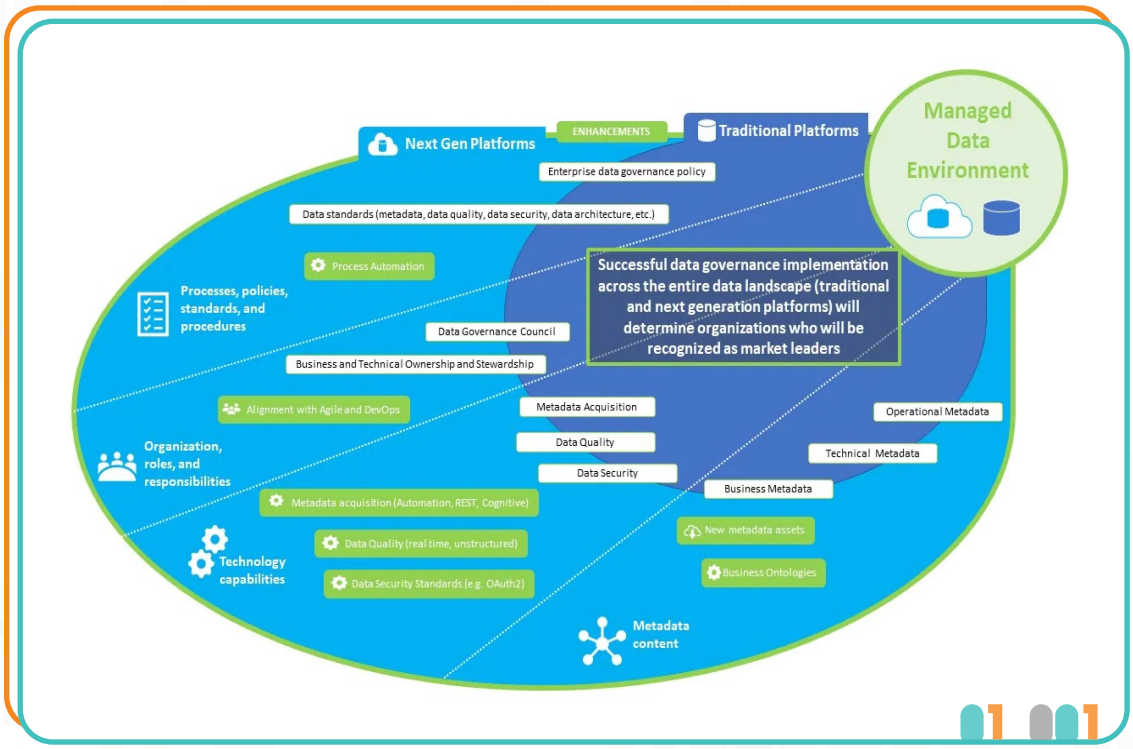



*Figure 4 - Collaborative - Data Governance Framework* *courtesy of Deloitte*

## 1.1 DATA GOVERNANCE IN THE MODERN DATA STACK.

As data teams have become more mainstream, the modern data stack has made it easier to ingest and transform enormous amounts of data, however the lack of data governance practices is one of the key barriers preventing data teams from being agile and driving forward with quick and intuitive response to provide a positive impact for the business.

This significant increase in data and therefore its potential has highlighted the importance of data governance, but it's also highlighted how underperforming the current data governance applications currently are. This dilemma has shown unequivocally the need for active data governance, which can be applied if ML capabilities are present in a governance application.

For the first time, the need for governance is being felt bottom-up by practitioners, instead of being enforced top-down due to regulation and security requirements. This bottom-up adoption is not only an opportunity for us to finally get data governance right, but also the only way data governance can succeed in 2024 and beyond. However, modern data governance for the modern data stack will look very different from its predecessor and this means that we'll have to change the way we approach data governance implementations.

The paradigm shifts that data governance needs today. Today's IT and business users will not tolerate long drawn-out governance projects that never quite meet expectations. Today governance projects need to be agile, be quick to implement, and provide immediate results that continue without manual intervention for the life of the product (i.e., ML capable and adaptive).

Data Mesh and Domain Driven design has become a real and positive architecture that is helping organisations to understand and harness their extraordinary growth in data. It may therefore make sense to also decentralise your approach to data governance. Let's look at these possibilities in the next section.

## 2.1 MOVE FROM CENTRALISED TO A HYBRID DATA ARCHITECTURE.

Data warehouse, data lakes – the strengths and weaknesses of a centralised approach

The technical infrastructure and organisational approaches for handling data are entering the next stage in their evolution. Since the 1990s, companies have held data for analytical purposes in central database platforms, known as data warehouses. The concept of 'data lakes' emerged as a response to the phenomenon of rapidly growing databases and mixed data constructs. This is designed to provide storage and processing for any data and is often operating in parallel to a data warehouse.

The growing number of promising use cases for decentralised data-driven solutions highlights the weaknesses of only a centralised approach. The operating model for a centralised approach cannot cope with the growing number of use cases across various domains. There are three principles for gaining greater benefits from data and analytics. To enable the creation of higher business value from data, data and analytics leaders aim to reduce bottlenecks and decentralise their data platform and teams.

**Four principles motivate the current evolution of the data & analytics area:**

*3.1.1.1 Scalable development of data applications*
More data & analytics use cases need to be rapidly developed and operationalised on a broader scale.

*3.1.1.2 High-quality and trustworthy data*
The productive usage of data solutions also requires high data quality – there is no room for "garbage in, garbage out" practices.

*3.1.1.3 Efficient administration of data and application cases*
The broad use of data and development of data-driven solutions across an organisation requires a higher level of guidelines and standards to support a hybrid model.

*3.1.1.4 Rapid dissemination of data that supports a common data model*
Improve the ease of integration of siloed data and its ability to be shared across the enterprise with inbuilt transformations at the domain level.

The data mesh and domain driven design architecture addresses these requirements with new architectural paradigms for data platforms. You probably know this term by now, even if you don't exactly know what it means. The idea of the "data mesh" came from two blogs written by Zhamak Dehghani in 2019, Director of Emerging Technologies at Thoughtworks and describes it as a type of data platform architecture that embraces the notion of data within the enterprise, by leveraging a domain-oriented, self-serve design.

Novon's experience has shown that there is a disconnect between the effort to implement and adopt the data mesh architecture principle and the effort to change the current working culture to support this new culture. We believe that in some cases that either the data mesh architecture is misunderstood or the effort to shift to a distributed architecture culturally is underestimated.

There is also a need for improvement in the self-service area. A large percentage of companies don't have specialised data analysis teams to serve the business, but unfortunately very few organisations offer a self-service data analytics solution as an alternative. Data mesh (and supporting governance), plus the right tool can accelerate analytics self-service adoption by an organisation.

Some data services make sense to leave centralised, this is why Novon recommends a hybrid architecture. Each organisation use case will be different so a careful review of requirements mapped to the two architecture types will be required before proceeding.

## 3.1 DATA DRIVEN GOVERNANCE SOLUTION ARCHITECTURE.

**Start with an architecture of a next Gen Stack that works for your organisation. It could be a mixture of**:

- Open-Source / best of breed products - Databricks, Snowflake, dbt, Apache Kafka, Spark, etc. or
- a Microsoft Azure collection plus additional vendors, or
- a Google Cloud solution plus additional vendors

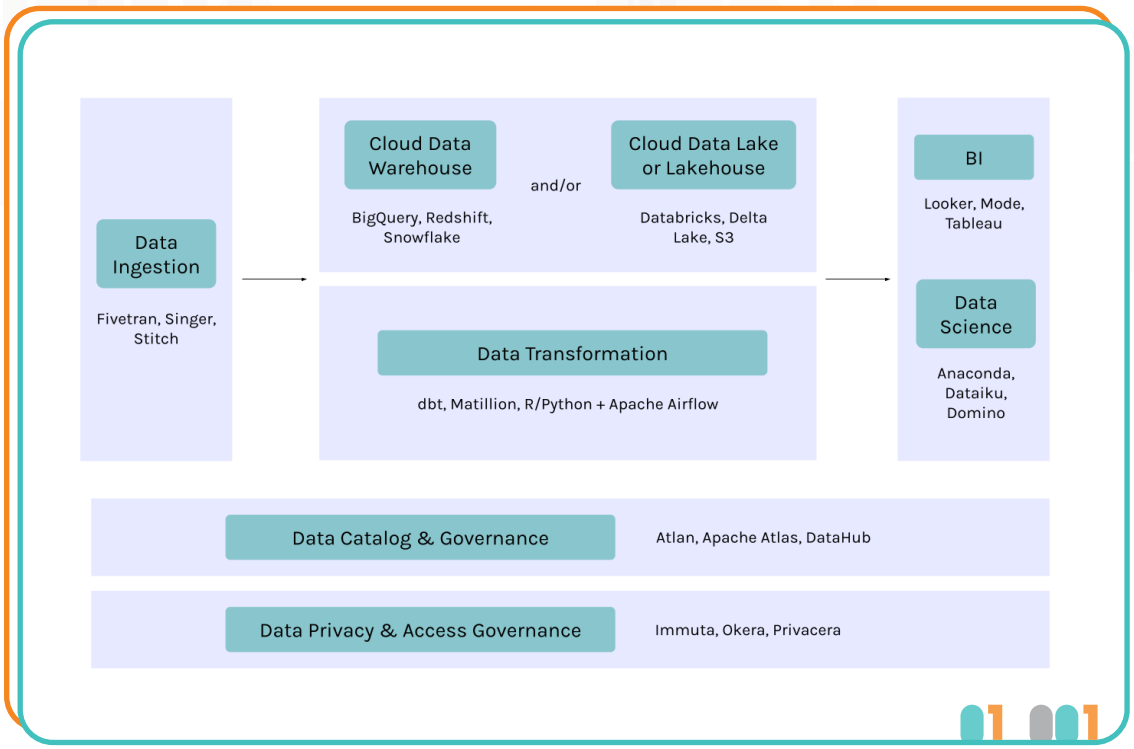



*Figure 5 - Open-Source / best of breed stack* *courtesy of Atlan*

## 3.2 THE RIGHT TOOL IS MORE IMPORTANT THAN EVER BEFORE.

Novon believes the answer lies in the tools selected by an organisation to perform data governance.

It is our belief that data is the key element in data governance, with the cultural, political, and commercial components of secondary importance. Having said that, the use of the data as determined by the business is also primary. From the 10 primary steps (see section 1.6 for the list) in data governance those that deal with the data and its purpose must be resolved and implemented first in the appropriate tool.

Data Strategy, Data Processes, (data issue tracking and resolution, data quality monitoring, data sharing, data lineage tracking, impact analysis, relationships, automated data quality testing, etc.), Data Policies, Data Standards, Data Rules, Data Security and Decision rights are important inputs to the total selection.

Enter the not so humble active metadata catalogue, that use AI / ML to automatically learn and continually update the data about the data, it then stores this data in an easily accessible data lake for use by everyone. Gartner has labelled this new tool category, Active Metadata. Throughout the industry this sector has now started to be called V3.0 data catalogues.

Novon believes this new data catalogue type can return data governance to a collaborative framework, which will drive the adoption and use of metadata to improve the timeliness of provisioning data. The new catalogue will add security features and attributes to your data not possible before and provide additional layers of privacy and security to your data where required.

# 4 When implemented correctly what should I expect?

In this section we discover and describe the key deliverables for V2.0 and V3.0 data catalogue solutions, however at this stage no vendors are mentioned, that will occur in the section five that follows.

What did happen, what is happening (real-time) and what could happen (predictive) are very difficult data problems to solve today. What is happening is the operational aspect of data governance that up until recently, there were little or no solutions for, now some vendors state they can deliver on this component. We will discuss this further in sections 5.

There are many tools on the market now that can help you with your data governance initiative. There are numerous products that hold and manage your data glossary, data catalogue, data dictionaries, but they are mostly passive or manual in their application. Up until very recently they haven't provided data lineage, data relationships, or the source applications for data. These would be generally labelled V2.0 data catalogue solutions. Today as we write this report, very few tools provide ML capability for active metadata management.

Regardless of these shortcomings, these earlier tools have proved popular and the number of vendors in the market has significantly increased over the last few years as organisations' data governance requirements gain momentum. Before we dive into today's metadata management world, let's take a quick look at the history of data catalogues, which can be broadly broken into three stages:

Data Catalogue 1.0 - Primarily built for IT users, acted as a data inventory.

Data Catalogue 2.0 - Data stewardship, provided tools for top-down data governance.

Data Catalogue 3.0 – Data collaboration, modern UX style, interactive real-time data points, typically using and supporting next gen technology.

## V2.0 DATA CATALOGUES

Let's look at the current V2.0 data catalogue offerings as a key baseline list to support your data governance projects, (unfortunately, most V2.0 catalogues are not ML capable) but most offer the following:

- May offer a cloud-based solution, may work across multiple domains and locations.
- Supports data mesh and domain driven design architectures.
- Data state – what format is the data in and how is it expressed.
- Data quality – cleansing, deduping, etc.
- Data Usage - data's importance, data security, and data confidentiality.
- Central data store for metadata.
- Can suggest data naming conventions based on existing column headings and hierarchy.
- Can suggest a data dictionary based on existing data.
- May provide data Lineage – only some vendor products.
- Can provide some relationships between data but not all – only some vendor products.
- Can provide most source data applications details.
- Data profiling.
- An easy-to-use old school structured UI / UX for a wide variety of users.

## V3.0 DATA CATALOGUES

The most significant difference between V2.0 and 3.0 catalogue solutions are the addition of ML capabilities. In other technology sectors like analytics, ML has been in use for 2 years or longer, providing predictive and prescriptive analytic views, often in real-time. Using ML, these mechanisms will help an organisation in pinpointing areas where data quality is compromised, and at the same time suggest how to rectify them.

Active metadata platforms act as two-way platforms, they not only bring metadata together into a single data repository, but also leverage reverse metadata to make metadata available in daily workflows when required. Their key offerings are:

- Cloud based solution – works across multiple domains and locations.
- Supports data mesh and domain driven design architectures.
- Data state – what format is the data in and how is it expressed.
- Data quality – cleansing, deduping, etc.
- Data Usage – can apply data importance, data security, and data confidentiality, retention policies, etc.
- Central data store for metadata – all data types.
- Can suggest data naming conventions based on existing column headings and hierarchy.
- Can suggest a data dictionary based on existing data.
- Using collaborative approaches, business context and the data dictionary can be updated automatically.
- Data Lineage.
- Relationships between data.
- Provide all source data applications details.
- ML algorithms and programable bots.
- Can provide in various formats - what happened, what is happening and what could happen?
- Data profiling.
- An easy to use and intuitive next gen UI / UX for a wide variety of users.
- Data and queries can be shared amongst users.
- Integration, both batch and rea-time is very easy with V3.0 tools, some are even supplied as part of the solution.
- Integrates seamlessly with next gen stack vendors (more then 3)

One additional item worth noting. To really give you governance flexibility and provide rapid response to day-to-day issues that arise, ensure you have three data model types, conceptual, logical and physical. All knowledgeable data practitioners know and use the logical and physical data models daily, so we won't dwell on those in this document. We will however discuss the conceptual data model and its uses in managing metadata.

Data relationships are key to dynamically modifying data governance configurations and constructs to meet evolving circumstances and business conditions. The conceptual data models retain these relationships, their definitions, and the semantics that separate them between departments or applications.

These conceptual data models assist with everything related to data governance, from rendering data access controls to facilitating lifecycle management necessities (like retention policies). Well defined conceptual data models are perhaps the starting point for adapting data governance protocols to meet arising situations. To that end, such models are comprised of a specific domain and support the data mesh architectural principles within them.

This is true for V2.0 and 3.0 data catalogue implementations, although V3.0 catalogues offer significantly more flexibility for real-time options on the fly.

# 5 Vendor brief and scorecard

Novon has taken every caution in preparing this information listed below, it is however, for the most part, based on website information provided by the vendor. Where Novon has worked directly with the vendor products and has gained additional insight we have added this information to improve the detail of the vendor in question.

The list of vendors is not exhaustive, but it does include vendors that either Novon has had some interaction with or one of their clients has worked with one of the vendor solutions below. These vendor solutions may also have been recognised in one of the industry research reviews, e.g., Forrester Wave.

We also wanted to include the two vendor perspectives, one for data governance and data catalogue vendors only. The second vendor perspective includes solutions where the governance and catalogue is part of the overall data management offering. Please see the Vendors included in Novon's web-based review below:

### Standalone data governance / data catalogue vendors.

- Alation
- Atlan
- Collibra
- Data World

### Data governance / data catalogue as part of total data management offer.

- Databricks - Unity
- Google GCP - BigQuery, data catalogue, ML, etc
- Informatica
- Microsoft Azure - Purview

Many of the vendors shown above have been acknowledged in Gartner's magic quadrant, or The Forrester Wave or IDC reports. Please see the Forrester graph on the following page as an example.

Today we're at an inflection point in metadata management. A shift from the slower semi-automated data catalogue V2.0 to the start of a new era with data catalogue V3.0 applications. As the jump from V1.0 to V2.0 was significant this will also be a fundamental shift in how we think about metadata and what we can do to harness its power to enhance our observations and actions for data governance.

Data catalogue 3.0 applications will not look and feel like their predecessors in the data catalogue 2.0 generation. Instead, data catalogue 3.0 apps will be built on the premise of embedded collaboration. Collaboration is seen as the key in today's modern workplace, borrowing principles from next gen applications like GitHub, Slack, Notion, and other tools that are commonplace today.

So, what should modern metadata solutions look like in today's modern data stack? How can basic data catalogues evolve into a powerful vehicle for data democratisation and governance? Why does metadata management need a paradigm shift to keep up with today's needs?

End-to-end data visibility, rather than piecemeal solutions. Data catalogue 2.0 era tools made significant strides in improving data discovery, however, most didn't give organisations a "single source of truth" for their data. Information about data assets is usually spread across different places, data lineage tools, data quality tools, data prep tools, and more. Data catalogue 3.0 apps will help teams finally achieve the holy grail, a single source of truth about every data asset in the organisation.

It's time for a modern metadata solution, one that is just as fast, flexible, and scalable as the rest of the modern data stack, which translates to a collaborative workspace for diverse data users.

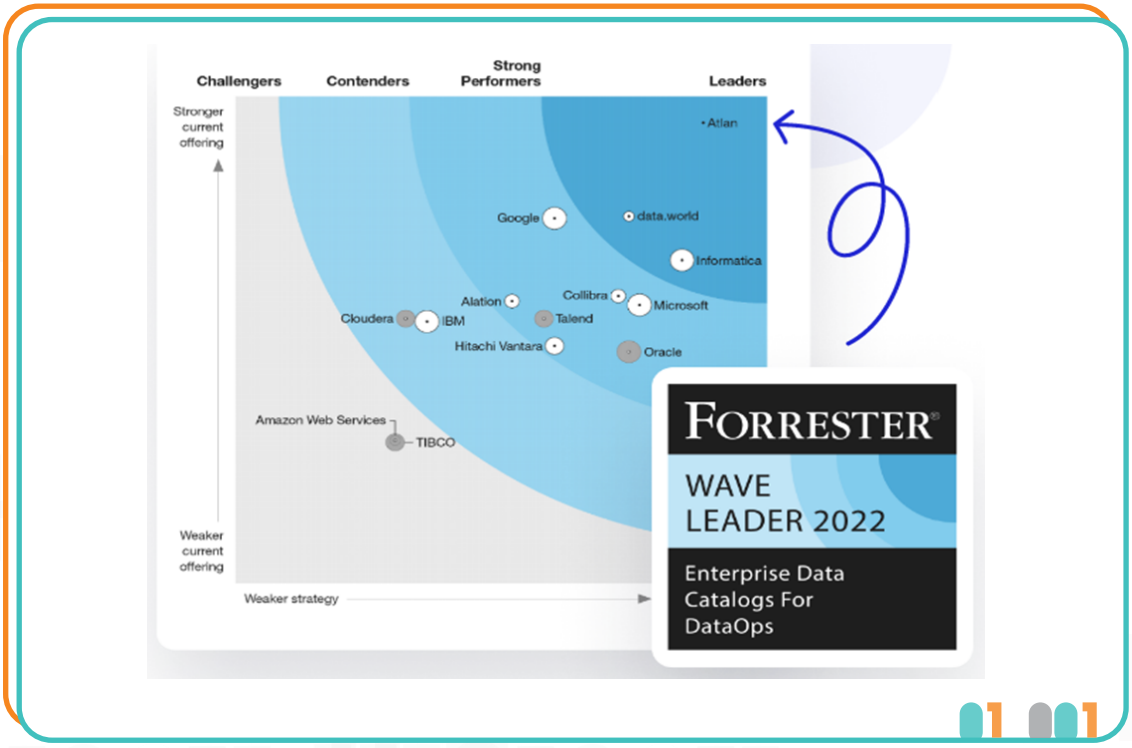



*Figure 6 - Forrester Wave - Data Catalogues*

## 5.1 SUMMARY - DATA GOVERNANCE / DATA CATALOGUE ONLY VENDORS.

| Feature List | V2.0 FEATURE | V3.0 FEATURE | COMPANY | COMPANY | COMPANY | COMPANY |
|---|---|---|---|---|---|---|
| | | | Alation | Atlan | Collibra | Data World |
| Cloud based solution. | X | X | X | X | X | X |
| Data state - what format is the data in and how is it expressed. | X | X | X | X | X | X |
| Data quality - cleansing, deduping, etc. | X | X | X | X | X | X |
| Data Usage - can apply data importance, data security, and data confidentiality, retention policies, etc. | X | X | X | X | X | X |
| Central data store for metadata - all data types. | X | X | X | X | X | X |
| Can suggest data naming conventions based on existing column headings and hierarchy. | X | X | X | X | X | X |
| Can suggest a data dictionary based on existing data. | X | X | X | X | X | X |
| Using collaborative approaches, business context and the data dictionary can be updated automatically. | | X | X | X | X | X |
| Data Lineage. | X | X | X | X | X | X |
| Relationships between data. | X | X | X | X | X | X |
| Provide all source data applications details. | X | X | X | X | X | X |
| ML algorithms and programmable bots. | | X | X | X | 2022 upgrade provides new features | X |
| Can provide in various formats - what happened, what is happening and what could happen? | | X | X | X | X | X |
| Data profiling. | | X | X | X | X | X |
| An easy to use and intuitive next gen UI / UX for a wide variety of users. | | X | X | X | X Very close | X |
| An easy to use old school structured UI / UX for a wide variety of users. | | | | | UI has been improved but may lack some next gen features | |
| Data and queries can be shared amongst users. | | X | X | X | X | X |
| Integration, both batch and real-time is very easy with V3.0 tools, some are even supplied as part of the solution. | | X | X | X | X | X |
| Integrates seamlessly with next gen stack vendors (more than 3) | | X | X | X | X | X |

*Highlighted features in orange, considered very important for modern active metadata management.*

## 5.2 SUMMARY - DATA GOVERNANCE / DATA CATALOGUE AS PART OF TOTAL DATA MANAGEMENT OFFER.

| Feature List | V2.0 FEATURE | V3.0 FEATURE | COMPANY | COMPANY | COMPANY | COMPANY |
|---|---|---|---|---|---|---|
| | | | Databricks | Google GCP | Informatica | Azure |
| Cloud based solution. | X | X | X | X | X | X |
| Data state - what format is the data in and how is it expressed. | X | X | X | X | X | X |
| Data quality - cleansing, deduping, etc. | X | X | X | X | X | X |
| Data Usage - can apply data importance, data security, and data confidentiality, retention policies, etc. | X | X | X | X | X | X |
| Central data store for metadata - all data types. | X | X | X | X | X | X |
| Can suggest data naming conventions based on existing column headings and hierarchy. | X | X | X | X | X | X |
| Can suggest a data dictionary based on existing data. | X | X | X | X | X | X |
| Using collaborative approaches, business context and the data dictionary can be updated automatically. | | X | First review suggests data steward approach | X | X | X |
| Data Lineage. | X | X | X | Only for Google data | X | Only for data passing through Azure data factory |
| Relationships between data. | X | X | X | X | X | X |
| Provide all source data applications details. | X | X | X | Only for Google data | X | Only for data passing through Azure data factory |
| ML algorithms and programmable bots. | | X | X | X | X | X |
| Can provide in various formats - what happened, what is happening and what could happen? | | X | X | X | X | X |
| Data profiling. | | X | X | X | X | X |
| An easy to use and intuitive next gen UI / UX for a wide variety of users. | | X | | X | There are varying opinions regarding the ease of use and intuitive feel, but overall a good UI | |
| An easy to use old school structured UI / UX for a wide variety of technical users. | | | X | | | In between old and new style |
| Data and queries can be shared amongst users. | | X | X | X | X | X |
| Integration, both batch and real-time is very easy with V3.0 tools, some are even supplied as part of the solution. | | X | X | Only for Google data | X | X |
| Integrates seamlessly with next gen stack vendors (more than 3) | | X | X | Doesn't support other data sources | X | X |

*Highlighted features in orange, considered very important for modern active metadata management.*

## 5.3 ALATION

### Summary

Alation is a standalone data governance provider established in June 2012. Their website states they are leaders in data intelligence, but its data intelligence for data governance.

Alation provides an open and intelligent platform that supports a wide variety of metadata management applications from search & discovery to data governance to digital transformation.

### Feature scorecard %

95%

### V2.0 (data steward driven) or V3.0 (collaboration first)

The launch of their first product was data steward driven, but because of when it was launched, it also included several V3.0 features, making it difficult to label V2.0 or 3.0. It's certainly a very strong contender in today's data catalogue offering.

### Modern UI / UX (collaborative bias) or old-style UI / UX (technical bias)

Modern UI / UX, but very structured, with some V2.0 data catalogue thinking embedded in the UX experience.

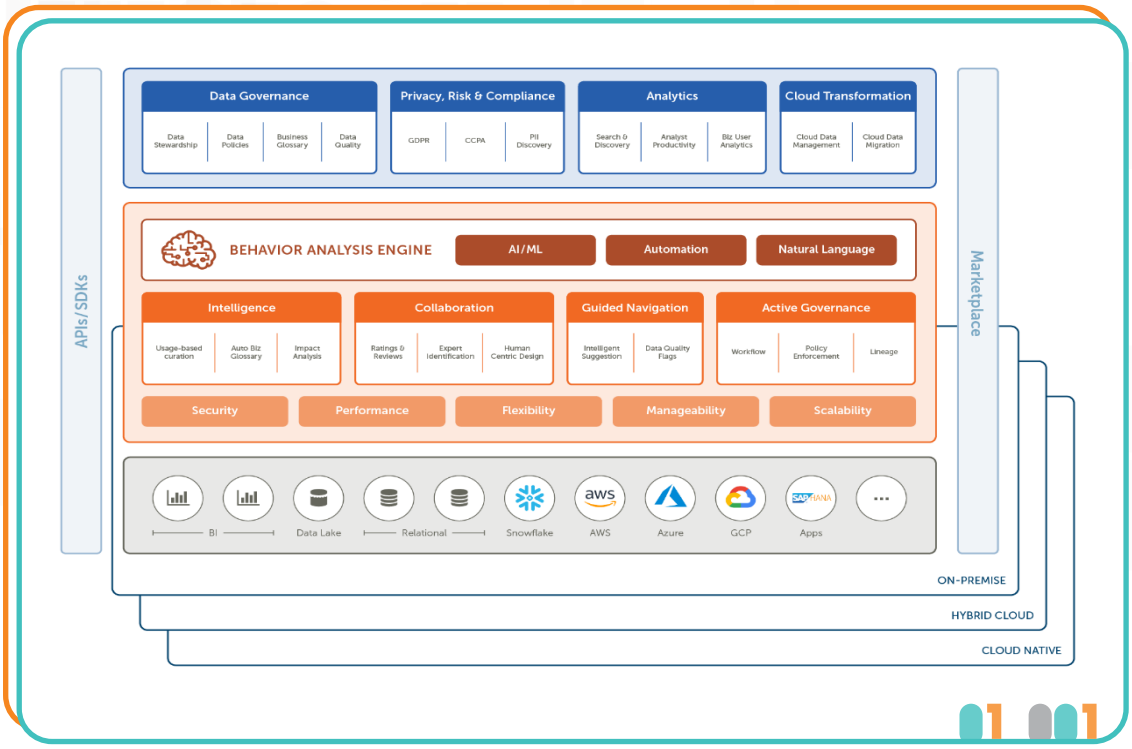



*Figure 7 - Logical architecture for Alation*

## 5.4 ATLAN

### Summary

Atlan is a standalone data governance provider that was established in 2019. They are evangelist about data and governance. The product is all about the next gen stack and the experience is based on collaboration across all aspects of the product. One of the co-founders coined the phrase data catalogue V3.0.

### Feature scorecard %

100%

### V2.0 (data steward driven) or V3.0 (collaboration first)

V3.0, collaboration first in all aspects of the product and its architecture, even their culture seems to be collaborative.

### Modern UI / UX (collaborative bias) or old-style UI / UX (technical bias)

A very modern UX, allowing for any user to participate in the data governance process. The UI is very intuitive and provides for the younger creative thinking approach more so than traditionally structured products.

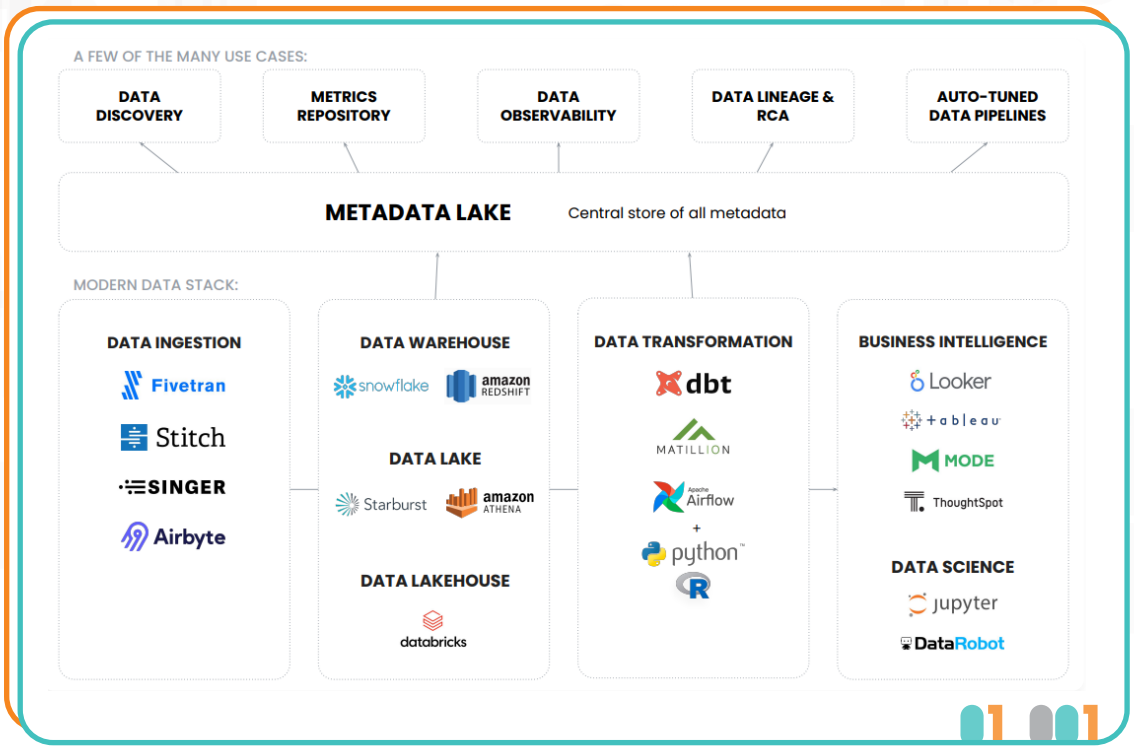



*Figure 8 - Conceptual architecture for Atlan*

## 5.5 COLLIBRA

### Summary

Started in June 2008 and was one of the first (if not the first) standalone data governance providers. Naturally being one of the first providers gave Collibra considerable market awareness and penetration, however it also meant significant changes were required to provide V3.0 features for their customers.

In early and late-2022 Collibra added a range of features inc. ML that has catapulted it back to its leadership status in data governance.

### Feature scorecard %

95% - requires additional study to confirm.

### V2.0 (data steward driven) or V3.0 (collaboration first)

The Collibra experience seems capable of providing for both engagement styles, however more time would be required with the product to determine its best approach.

### Modern UI / UX (collaborative bias) or old-style UI / UX (technical bias)

A significant step forward after the enhancements in 2022 but more time is required with Collibra to be able to determine its pedigree, for now Novon would rate the UI modern.

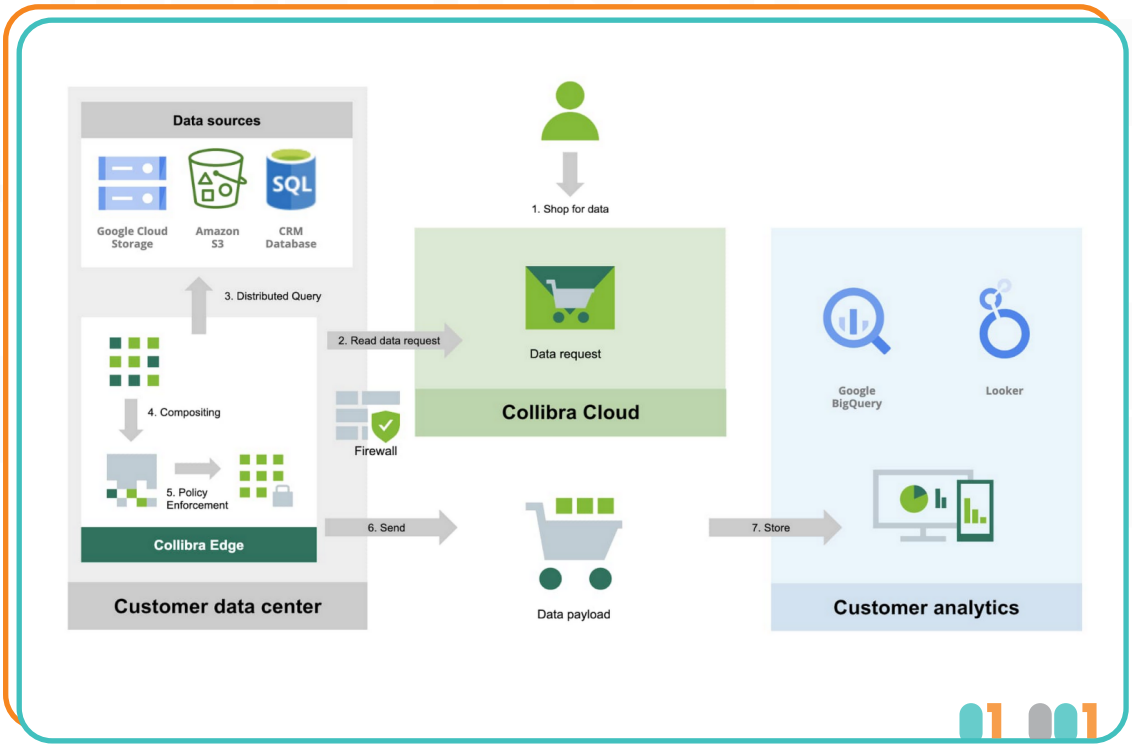



*Figure 9 - Conceptual architecture for Collibra*

## 5.6 DATA.WORLD

### Summary

Started sometime late 2015, early 2016. There's not a lot that we know about data.world, only what is on their website. They seem to work mostly in the US. They are definitely a data evangelist like most tier 1 data governance / data catalogue providers.

### Feature scorecard %

98% - to be confirmed.

### V2.0 (data steward driven) or V3.0 (collaboration first)

Like Alation, the launch of their first product was data steward driven, but because of when it was launched, it also included several V3.0 features, making it difficult to label V2.0 or 3.0. It's certainly a very strong contender in today's data catalogue offering.

### Modern UI / UX (collaborative bias) or old-style UI / UX (technical bias)

Novon has not been able to secure a demonstration at this time of the review, but looking through their website it looks to be a very modern UI / UX experience. What we cannot gather from the information available is if its data steward or collaboration driven.

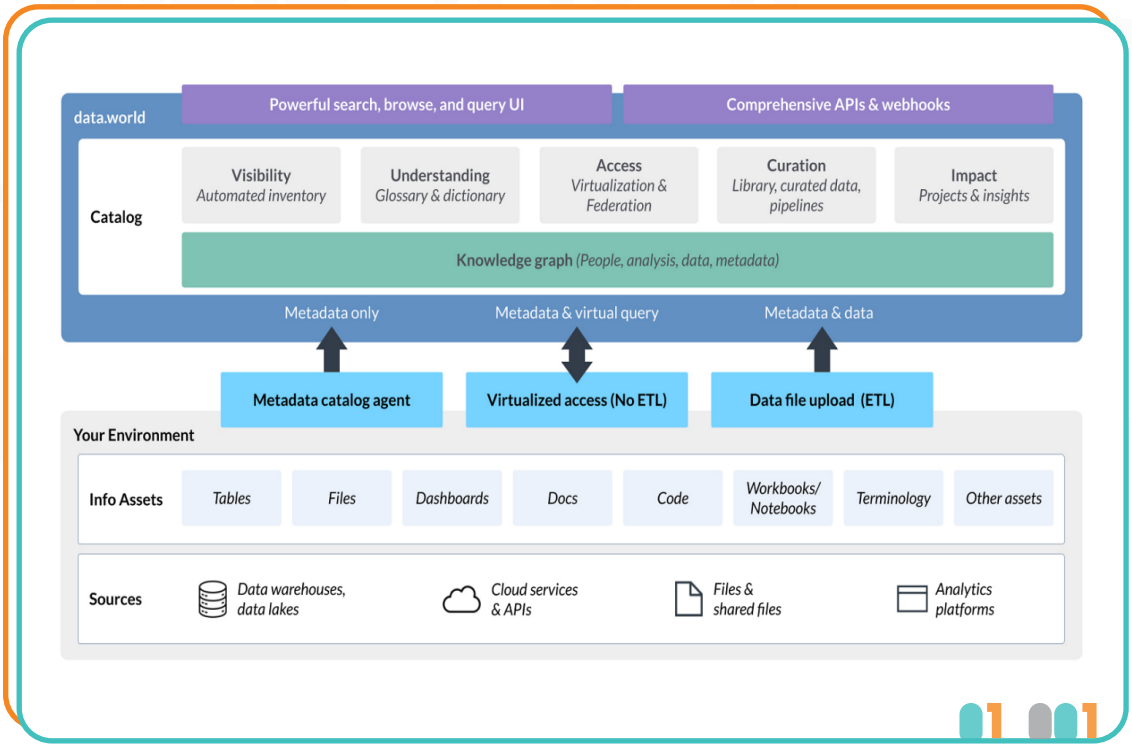



*Figure 10 - Logical architecture for Data.world*

## 5.7 DATABRICKS

### Summary

Databricks started in 2013, with their key product being Apache Spark. In May of 2021 Unity their data catalogue product was launched. Unity is feature rich and looks to target the technical teams in an organisation. This is not unlike GCP which does the same.

### Feature scorecard %

95%

### V2.0 (data steward driven) or V3.0 (collaboration first)

Based on our current review the Unity data catalogue looks to be data steward driven, although it could well be perceived as collaborative by users. We cannot say for sure until we complete further investigations.

### Modern UI / UX (collaborative bias) or old-style UI / UX (technical bias)

While the UI / UX experience is modern and somewhat collaborative, it seems to targets the technical part of an organisation much like a database driven organisation would. For this reason, we would mark the UI older in style. However, we wish to point out that this approach does not reduce its effectiveness rather it narrows its audience.

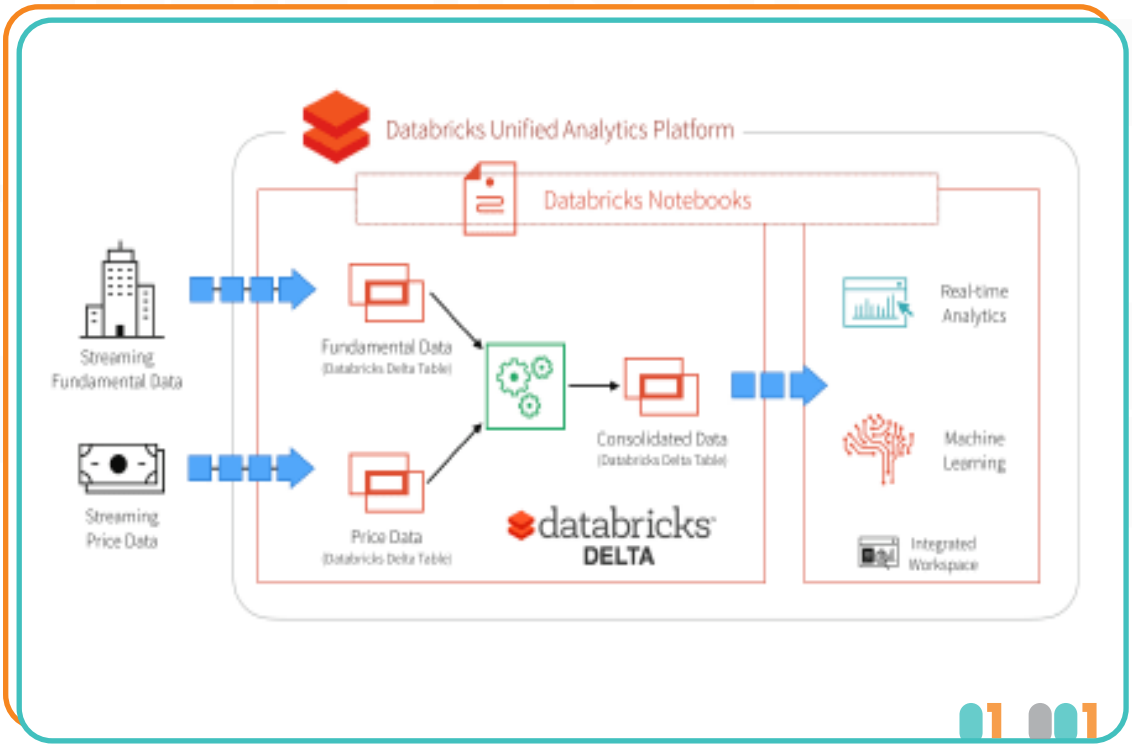



*Figure 11 - Conceptual architecture for Databricks*

## 5.8 GOOGLE BIGQUERY (GBQ)

### Summary

A recent addition to the GBQ stable, Dataplex offers a very competent data catalogue solution, however it does not (at this stage) support other data sources or on-prem data sources it is still a very strong Google only data catalogue. The lack of other data sources support is a significant gap in this product offering and downgrades it rating considerably.

### Feature scorecard %

90%

### V2.0 (data steward driven) or V3.0 (collaboration first)

Dataplex is an intelligent data fabric that helps you combine distributed data and automate data management. Data governance can be added across that data to power analytics at scale.

### Modern UI / UX (collaborative bias) or old-style UI / UX (technical bias)

As you would expect from GCP the UI / UX experience is modern and collaborative.


*Figure 12 - Conceptual architecture for GCP - Dataplex*

## 5.9 INFORMATICA

### Summary

Informatica released their current native cloud data catalogue version in July 2021. This new comprehensive solution brings together data cataloguing, data quality, data and AI governance capabilities with unified metadata-driven intelligence natively in the cloud. Cloud Data Governance and Catalogue provides the industry's first integrated solution for governing both the AI models and the data that feeds the models.

### Feature scorecard %

95% - It would be 100% except for the bias in the UX to support technically capable people.

### V2.0 (data steward driven) or V3.0 (collaboration first)

Like a few of the new versions released but with a strong history in data management the solution tends to lean towards data stewardship first and collaboration second.

### Modern UI / UX (collaborative bias) or old-style UI / UX (technical bias)

It is a modern UI / UX, however it does lean toward a technical user experience. There is allowance for business users but Novon has not had time to properly evaluate these features.

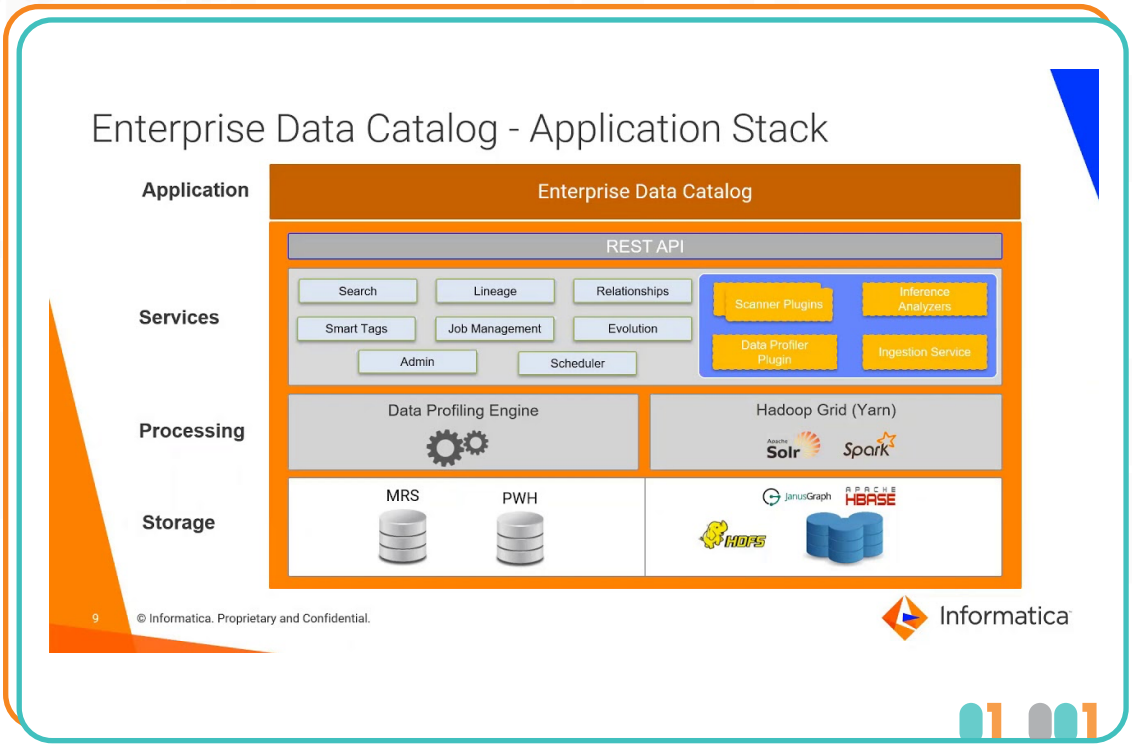



*Figure 13 - Conceptual architecture for GCP - Dataplex*

## 5.10 MICROSOFT AZURE

### Summary

In September 2021 Microsoft released their Azure metadata catalogue - Azure Purview, (previously Microsoft Purview). Purview is designed to enable an organisation gain visibility into assets across their entire data estate. It will leverage visibility to manage end-to-end data risks and regulatory compliance and will generally govern, protect, and manage data in a new, more comprehensive, and simpler way. Unfortunately, Purview only supports data lineage for Azure data factory. Novon expect Microsoft to address this fairly quickly.

### Feature scorecard %

90% - older style UI that is technically focused and data lineage not available for non-Azure data factory integrated data.

### V2.0 (data steward driven) or V3.0 (collaboration first)

The tool seems to be a blended solution, providing for both data stewards and some experienced business users. The most important point however is that the tool seems to be targeting the technical people within an organisation, not everyday business users.

### Modern UI / UX (collaborative bias) or old-style UI / UX (technical bias)

The UX experience is a blended version of collaboration and technical. A business user would need to be quite comfortable with technical data terms and UI activities to use Purview effectively. In Novon's experience only finance or engineering departments could qualify.

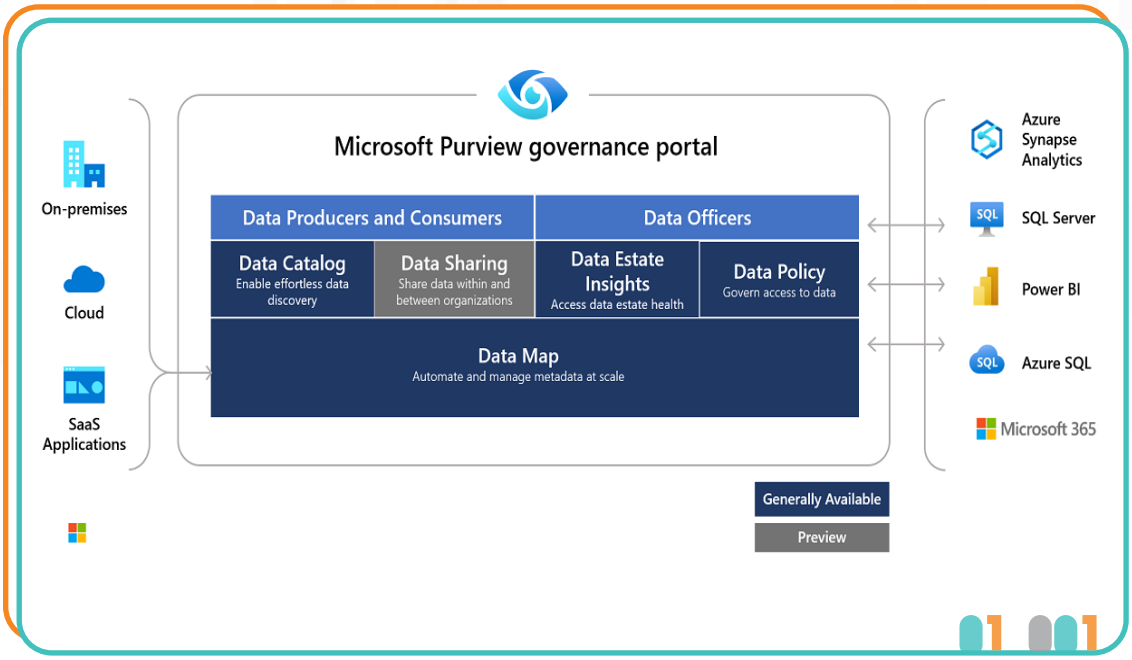



*Figure 14 - Conceptual architecture for Azure Preview*