Bloor

White Paper

White Paper by Bloor

Author **Andy Hayler**

February 2025

Data Governance in the age of AI

In 2025 data governance has implications for data fabric, compliance, data standardization and beyond.

Introduction/background

Organizations today face a range of challenges in managing their data assets. Data volumes are exploding, the number of data sources is increasing, and all this is occurring while there is a gradual move of more and more applications to the cloud, whether public cloud(s), private cloud or a hybrid model alongside on-premises applications that are not going away any time soon. The exact number of data sources varies by enterprise, but one survey puts it at 400 on average, with large organizations having to cope with over a thousand data sources. 64% of organizations have to deal with at least a petabyte of data, and 41% have to handle over 500 petabytes in 2024. This rate is growing by at least 22% annually, with some estimates much higher than this. In all, 78% of enterprises have a multi-cloud or hybrid environment.

To deal with all this complexity, a range of technologies have been deployed. In this whitepaper, we will discuss the current trends, what they have to offer, and some of the issues that arise.

> “Data volumes are exploding, the number of data sources is increasing, and all this is occurring while there is a gradual move of more and more applications to the cloud, whether public cloud(s), private cloud or a hybrid model alongside on-premises applications that are not going away any time soon.”

Data fabric

With hundreds of data sources, it is hard for an enterprise to answer seemingly simple questions like *"What is my most profitable product and customer?"* since there are multiple versions of customer and product scattered amongst the data sources. A Deloitte Digital survey found an average enterprise has 17 different systems housing customer data, a number that rises to over 50 systems for 62% of US retailers. The traditional approach has been to improve the quality of such data by deciding on which sources are the most trustworthy or recently updated, and coming up with a single "golden record" for customer and product data based on business rules, human judgement, or a combination of both. Hubs of such high-quality master data can then be used to populate the dimensions of a data warehouse, giving context to each business transaction of the company. However, this involves moving a lot of data around, typically in overnight batch routines. The data warehouse is therefore not completely up to date at any one moment, and it is further hampered when changes happen to the structure of the underlying source systems, say, in the event of a reorganization or acquisition. When this happens, the data warehouse will be slightly out of line with the source systems until the warehouse structure can be modified, something that typically takes weeks to do.

A different approach is what is called a data fabric (or a related approach called data mesh). In the case of a data fabric the source data is left in place, and a business-oriented semantic layer is developed that sits above the data sources, mapping business terms like "customer" onto underlying sources, shielding business users from physical structures like table and column names. Business queries are presented through a "knowledge graph", a kind of treasure map of data assets that uses terminology that is meaningful to the end users. The users can essentially shop around for the data that they need by exploring this visual description of data assets, rather like they might shop for gifts in a marketplace. This has the advantage of removing the need for mass data movement, with queries satisfied by direct access to the underlying sources, translated from business queries into underlying database queries by the semantic layer of metadata. The knowledge graph exposes data assets to users in their own terms, much improving visibility of data, which in itself improves compliance. In some implementations, users can score and comment on the data assets rather like they might leave reviews of products that they bought online, which in turn raises the visibility of any issues with the data and encourages data quality to be improved.

The data fabric is not a panacea, as the hard work of integrating multiple candidate data sources still has to be done somewhere, in this case in the semantic layer. There are also performance issues to be considered for the generated queries, as these may involve distributed queries to various source systems. Nonetheless, if this can be achieved, then end users will be able to query data in real-time without having to learn SQL or be aware of the underlying physical structure of data. Data mesh, incidentally, has a similar philosophy of leaving the source data in place, but with data mesh the data governance is devolved to business units rather than being centralized.

"Business queries are presented through a "knowledge graph", a kind of treasure map of data assets that uses terminology that is meaningful to the end users."

Data Standardization

As mentioned, one key issue for any large organization is the diversity of data sources and lack of agreement about common terminology. Different business units and company subsidiaries in global companies tend to develop their own jargon and terminology. Everyone can agree that having a single common definition of something like "net profit" or "inventory turnover" or "accounting period" is a good idea, but getting everyone in the company to agree to change theirs to a new standard is another thing. The author of this paper recalls facilitating a meeting of a global company where six different procurement groups around the world each presented to each other at a workshop the necessity of coming up with a single standard for procurement terminology, only to spend the bulk of each presentation explaining why their particular set was best and why everyone else in the world should change to theirs. Even seemingly rock-solid terms like "cost of good sold", which is a well-defined accounting term, can end up being interpreted differently by different business units that decide what to include or not to include. Precision of definition is vital. Business glossary entries should not just have a formal definition, but explanatory text and examples if needed.

There is no shortcut to the process of hashing out such issues and agreeing on who owns what data asset (such as "product" and its classification hierarchy), who ultimately can decide on the master version of that terminology for the company and how to resolve disputes about it between business units. Named individuals need to be assigned and authority granted to them, and senior management needs to provide backing to the efforts to change the way things are currently done and to move to new standard terminology. If this data governance process can be successfully implemented then the common terminology can be captured in a business glossary that is available for all to see. This business glossary becomes part of a broader data catalog that documents the data assets of an enterprise, including who owns them.

Some types of data do not change very much. Whereas a list of customers may change every day or even every minute, some data lists of currencies, country codes, units of measure, calendars, conversion rates etc either never change or do so very rarely. A new country like South Sudan might pop up from time to time, or a country may split into others, such as when the Soviet Union split up, but such events are rare. These types of "reference data" are good candidates to be agreed on globally with an organization, and a master list of things like currency codes can be referred to by the business glossary or data catalog.

Standardizing definitions and data across an enterprise brings many benefits. People spend less time arguing about whose version of data is correct and more time on addressing business problems. Trust in data improves and the quality of data is ensured. Communication improves and new employees can be onboarded more efficiently as a common business language is adopted. Common definitions reduce the risk of confusion when a company undergoes an audit or needs to provide regulatory reports to the government. As a further benefit, application integration efforts are eased, bringing tangible cost savings.

An important element of data governance is not regarding it as a one-off exercise: it is a journey rather than a destination. Companies need to set up ongoing committees and resources to maintain the standards once they have been established, to deal with the inevitable changes that will occur from time to time and to resolve debates and arguments about these standards.

> "An important element of data governance is not regarding it as a one-off exercise: it is a journey rather than a destination."

Operationalization

Getting high-level agreement to standardize and document business terms in an enterprise is one thing: implementing it is another. It is common to assign responsibilities to a centrally resourced team that supports assigned data stewards embedded within business units. Overseeing things is usually a steering committee of senior business executives with enough political clout to resolve disputes between different lines of business if issues cannot be dealt with at the working level and need to be escalated.

The job of the data steward is to ensure the accuracy of data within their area of responsibility, enforce data governance rules and maintain data standards. This may extend to ensuring the security and privacy of data, and complying with data retention policies. In cases where there are multiple source systems holding data on a subject, such as about customers, products or assets, then the data steward will be responsible for deciding on the rules that determine which systems are the most trusted in the case of overlapping or duplicated data. They may need to review data in situations where data records are flagged as potential duplicates, either by colleagues or by data quality software. The goal of a data steward is to ensure that the data within their remit is of the best quality and can be relied on. Trusted versions of data records will be used by queries that are handled and generated by the semantic layer in the case of a data fabric or mesh architecture.

An organization needs to document its data assets and set up business rules that ensure its accuracy as far as possible. The scope of this may involve structured data in relational databases but also document and image libraries, as well as reference data. Enterprises need to keep track of the versions of data too, a concept known as data lineage. For example, there may be a new version of a customer classification hierarchy that becomes operational on a certain date. Business users need to know which version of data they are dealing with, and potentially be able to look at prior versions in some situations, or for audit or compliance purposes. Knowing the value of sales in a certain region or the marketing costs in another is one thing, but for compliance reasons this information may be needed across a period of time, even though there may have been changes to classifications over the period in question. Clearly, such data lineage issues are ideally handled by specialist software that is designed for the purpose. As well as databases and documents, there may be other items that need to be managed, such as AI models. These models are just another kind of data to be dealt with, and they themselves have versions, performance parameters and characteristics, such as what data they were trained on, which may include company-specific data where retrieval-augmented generation has been used.

Technology can help in the process of building up a map of data assets. Some tools can scan the database catalogs and metadata in an enterprise and come up with candidate lists of data assets, including showing the relationships between data types. For example, patients may be assigned to a particular doctor, specific customers to a certain sales representative, or retail stores to a particular organizational region. There may need to be manual adjustment, but automated scans of this type can save time and reduce the effort in what may seem initially like a mountainous task. A 2024 survey by Experian of 250 data governance decision makers found that 84% of survey respondents felt that data governance was the backbone of effective data management, yet just 15% felt that their current data governance efforts were meeting expectations. This indicates the difficulty in moving from high level aspirations to an operational, working system.

> “The goal of a data steward is to ensure that the data within their remit is of the best quality and can be relied on.”

Policy-driven compliance enforcement

A well-implemented data catalog and business glossary can greatly help implement corporate data policies. These may include how different classifications of data need to be handled, such as personally identifiable data needing to be masked. It may extend to defining retention periods for data and requirements for data storage and deletion in some cases. Other data policies will cover data access and security, and the management of data across its lifecycle, from creation to archiving and disposal. There may even be data ethics policies that protect the rights of data subjects and promote fairness and accountability, as well as more mundane but important issues such as ensuring the integrity and accuracy of data, its completeness, validity and timeliness.

All these data policies need to be monitored once they are made operational. It is important that the level of compliance with the policies is monitored, exceptions reported and recorded and the level of quality of data be reported over time. Good data catalog and data quality software will provide comprehensive reporting and analysis capability, so for example, you can see how the quality of each data asset is holding up over time, and whether remedial action needs to be taken if the levels drop below a threshold set by a data policy. Compliance with data policies is not just a good thing in itself. It can avoid exposure of the enterprise to legal and compliance risks in some cases. Different jurisdictions have differing regulations affecting data, and these can affect an enterprise in any of the territories where they operate, not just where they are headquartered.

There are numerous examples of large fines being levied by regulators for failures in data compliance, such as breaches of the General Data Protection Regulation (GDPR) in Europe. Fines have on several occasions been over €100 million in value and have affected some well-known companies (Amazon was fined €746 million for GDPR violations in 2021 for example, and META was fined €1.2 billion in 2023 under the same legislation). Investing in a properly resourced data governance program can have a positive return on investment, if only to help avoid such regulatory issues.

Another area where compliance may be important to consider is environmental, social and governance (ESG) reporting. Approaches here vary by jurisdiction, but the EU has regulations around ESG, and many US states have put in place rules around ESG investments, which itself has a knock-on effect on companies. Yet another area to consider is that of data sovereignty, the principle that data is subject to the laws and regulations of the country or region where it is stored, collected, or processed. This is particularly relevant today given the move towards cloud processing, where the location of the servers holding company data may not be in direct control of the company in the way that a data centre is. Legislation applies here too, not just the EU's GDPR rules but also the California Consumer Privacy Act. There are further laws affecting data sovereignty, from Australia to Saudi Arabia and China. This maze of legislation affecting data needs to be navigated and managed.

An effective data governance program can help with compliance as it will improve the quality of data in an organization. Data quality improvements can bring about improved customer satisfaction, better-targeted spending and reduced operational costs, and there are many well-documented examples of these types of benefits across a range of industries.

> "Data quality improvements can bring about improved customer satisfaction, better-targeted spending and reduced operational costs, and there are many well-documented examples of these types of benefits across a range of industries."

Conclusion

A modern enterprise faces significant challenges in managing its data, ranging from issues of data complexity and data quality, to handling sensitive data and compliance with regulators in different jurisdictions. Whether an enterprise is using a data warehouse or data lake architecture, or a data fabric or data mesh approach, a comprehensive data catalog and business glossary can support data governance initiatives and help to ensure regulator compliance. The same data infrastructure can help improve data quality and encourage a common data language in an enterprise through a well-implemented business glossary. These initiatives can ease the effective sharing of data within an enterprise and lead to better and more targeted business decision-making.

About the author

ANDY HAYLER
Practice Leader: Data as an Asset

Andy is an established software industry authority, an independent strategy consultant advising corporations, venture capital firms and software companies. He is the founder of Kalido, which under his leadership was the fastest growing business intelligence vendor in the world in 2001. Kalido was recognised as an innovator in data warehousing, and then launched arguably the first true master data management product, a market which at the time did not exist but is now a well recognised and fast growing industry. Andy was the only European named in Red Herring's "Top 10 Innovators of 2002". He was a pioneer in blogging with his award winning "Andy On Enterprise Software" blog.

Andy started his career with Esso, working in a number of technology roles before moving to Shell. He was Technology Planning Manager of Shell UK, then Principal Technology Consultant for Shell International. He later established a global information management consultancy, which under his leadership grew to 300 staff.

Bloor overview

Technology is enabling rapid business evolution. The opportunities are immense but if you do not adapt then you will not survive. So in the age of Mutable business Evolution is Essential to your success.

We'll show you the future and help you deliver it.

Bloor brings fresh technological thinking to help you navigate complex business situations, converting challenges into new opportunities for real growth, profitability and impact.

We provide actionable strategic insight through our innovative independent technology research, advisory and consulting services. We assist companies throughout their transformation journeys to stay relevant, bringing fresh thinking to complex business situations and turning challenges into new opportunities for real growth and profitability.

For over 25 years, Bloor has assisted companies to intelligently evolve: by embracing technology to adjust their strategies and achieve the best possible outcomes. At Bloor, we will help you challenge assumptions to consistently improve and succeed.

Copyright and disclaimer

Bloor Research International Ltd

20–22 Wenlock Road, London N1 7GU, United Kingdom

+44 (0)1494 291 992

info@Bloorresearch.com

www.Bloorresearch.com