MarketUpdate

Less data drama – a guide to data governance

What is it?

Data governance is the set of processes in an organisation that support high-quality data. These are the policies, standards and procedures that ensure an organisation can trust its data, including business ownership. There are many vendors in the space, with the market size depending on who exactly which ones you include. However, in 2024 the data governance software market was around $4 billion in size and growing at around 20%, faster than the 12% of the overall enterprise software industry.

What does it do?

Organisations have numerous data assets that they need to manage, including their databases of customers, products, assets, employees, contracts and transactions. Many of these data types span organisational boundaries, and data regarding things like customers and products are frequently duplicated in multiple systems. Even basic questions like which are our most profitable customers and products are tricky to answer if there are no standard definitions for classifications of products and cost allocations, and competing versions of key data. The idea of data governance is to assign business ownership of the core data assets of a company so that any conflicting definitions or overlaps can be resolved once and for all. Typically, there may be a data governance council with executive sponsorship, supported by a central dedicated data governance team as well as data stewards in the devolved business units. The idea of data governance is to use these resources to impose standards and policies across the organisation, and ensure data quality and regulatory compliance. The role of data governance may extend to data retention, data privacy and data security.

A software industry has sprung up to support these processes, and typically has at its heart a data catalog. This lists data assets and who owns them, providing definitions of business terms like *"net profit"* or *"gross margin"* in a business glossary. The catalog contains data policies and usually has features to report on compliance with regulations like GDPR or HIPAA. A data catalog will usually also include data lineage, showing the journey of data from its source system to its destination, for example from a data entry system through to a data lake or data warehouse, including any transformations that may occur along that journey. The data lineage provides an audit trail of data, showing changes and in this way supporting compliance reporting; it is like a family tree for data. Some vendors specialise in data security management and policy enforcement and do not attempt to handle all forms of business metadata or general data quality. This report does not include such security-oriented vendors. A data governance solution may go beyond the core catalog and include workflow support for data stewards, data discovery and data quality components. ▶

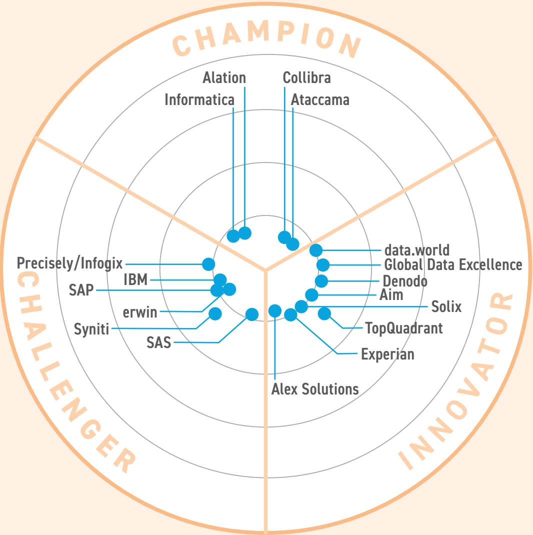

Figure 1

The highest scoring companies are nearest the centre. The analyst then defines a benchmark score for a domain leading company from their overall ratings and all those above that are in the champions segment. Those that remain are placed in the Innovator or Challenger segments, depending on their innovation score. The exact position in each segment is calculated based on their combined innovation and overall score.

Who should care?

Most organisations, whatever the industry and whether in the private or public sector, are increasingly dependent on data. Some companies deal entirely with digital assets, for example, streaming video services or banks without high street branches. Managing this data coherently is a challenge, with numerous surveys showing that only about a third of executives have a high degree of trust in their own data. That is an obvious problem, and in heavily regulated industries like pharmaceuticals and finance, an inability to demonstrate strong control of data can result in heavy fines in many countries. Leaving the quality of data to the IT department has been shown to be a recipe for trouble, as most IT departments lack the political clout to actually get business units to change their practices. The idea of data governance is to put control and ownership of data firmly with the business, and have assigned staff that can handle disagreements over customer classifications or product hierarchies or groupings of assets.

Although policies and standards can be defined without specific software, it is easier to actually implement and monitor policies with the help of dedicated data governance software. This imperative has seen the data governance software industry grow rapidly since the early 2000s into a large industry in its own right. Organisations large and small need to be aware of data regulations and need to be able to get a clear picture of business operations across the whole enterprise. This can only be done if there is consistent data, and data governance is a key building block to attaining that goal.

Emerging trends?

Data governance is affected by the same secular trends as other software, including the increasing shift to cloud computing, the rise of artificial intelligence, greater data democratisation, increased automation and new software architectures like data mesh and data fabric. Data governance solutions need to offer a range of deployment options, from public cloud to private cloud to on-premise and a hybrid of these. The data catalog needs to be able to ingest data from a wide range of data sources that reside on these various platforms, so increasingly needs real-time connectors to the various leading application platforms as well as a range of databases, document sources and more. This landscape may be represented by a "knowledge graph", a semantic network that shows interconnections and relationships between data entities.

The tsunami of investment in artificial intelligence (AI) affects data governance in several ways. For one thing the catalog needs to actually regard the AI models as data assets in themselves. These large language models have characteristics that should be tracked, including their usage, their accuracy, who has access to them, how often they are updated, resource utilisation etc. They are just another form of data asset to be managed. The data governance software can itself employ artificial intelligence in various forms. Machine learning has long been used in merge matching of data records, for example. Generative AI can be used to help populate business glossaries with descriptions that may otherwise be cumbersome and slow to do by hand, though this is best done under human supervision and review given this technology's tendency to "hallucinate". AI can also help discover and classify data assets and detect anomalies in the data. This is increasingly useful as the sheer range of different data sources increases: it is no longer enough to manage structured data in corporate databases. Depending on the industry, you also have to deal with documents, image libraries, multi-media assets like promotional videos, websites, sensor data and mobile device data, amongst others. A further twist is that companies are increasingly choosing to augment large language models with company-specific data, a technique called retrieval augmented generation. Such models are highly dependent on the quality of the data that they are trained on, so it is vital that data quality be well understood in deciding which data assets to use to train the AI models.mod

Successful data governance means engaging with business users, and modern software offers a consumer-like experience, displaying data assets rather like you might display goods in an on-line store, allowing business users to add comments and score the quality of data, just as they might review a movie on Netflix. The rise of the "data mesh" as an architecture implies decentralising the ownership of data throughout a business, and focusing on the creation of data assets, so is very much in tune with the core ideas of data governance.

Regulation continues to be a concern for many industries in most countries. Companies these days have to be aware of environmental, social and governance concerns of investors and the public in addition to direct government regulation of data such as that of the California Consumer Privacy Act, UK Data Protection Act, China's Personal Information Protection Law and many more, depending on the jurisdictions in which you operate.

Vendor Landscape

The data governance landscape is quite dynamic. Large vendors increasingly offer a broad data management platform in which data governance is just part. For example, as well as a data catalog they may offer data quality tools, metadata management, master data management, analytics, discovery and monitoring tools. Newer tools are appearing, often with a significant AI component. These tools often offer greater automation in terms of data discovery and classification than traditional products, though inevitably, larger vendors respond to such innovations with their own alternative approaches.

The current leading players in the market are Collibra, Alation, Ataccama and Informatica. These vendors have broad functionality across data governance and significant installed bases of customers. There is a further set of vendors who have solid offerings but less penetration in terms of dedicated data goverance customers including IBM, SAS, Quest (Erwin), SAP, Syniti and Precisely. There is a further cluster of vendors that are typically smaller in scale but have unusual or interesting innovative features: AlexSolutions, Denodo, Experian, Solix, Global Data Excellence, Top Quadrant and Aim Ltd. ▶

In general, many data governance vendors have been moving quickly to add data lineage capabilities to their products even since IBM acquired Manta. Several other vendors used to rely on Manta for their data lineage, which was fine as an independent company with scope limited to lineage, but not when acquired by a major competitor such as IBM. Several vendors have introduced major software releases in 2024, including Ataccama and erwin.

End users need to carefully evaluate their needs and decide whether they require the state of the art in specific areas that are important to them, or whether they prefer a broader solution that is not necessarily optimal in each area but is within a single solution, at least superficially (some large vendors have increased their scope through acquiring other tools, and their level of integration may be paper thin).

Conclusion

Data governance continues to grow rapidly as more and more organisations realise that the management of their core data assets is a key competitive advantage in an increasingly digital world. In some cases, data governance is imposed by regulation, but it also can open up opportunities. Companies that have high-quality data and a clear picture of their global business operations can more rapidly respond to changing circumstances and more quickly take advantage of opportunities than those that do not.

It is recommended that you carry out a detailed investigation and evaluation using your own data rather than relying on vendor demonstrations (which are carefully designed to always work perfectly) based on the specific needs of your own organisation. You may find it useful to engage a third-party expert experienced in evaluating these technologies to help you.

We have a designated webpage for this topic so for the latest research and commentary please visit HERE.