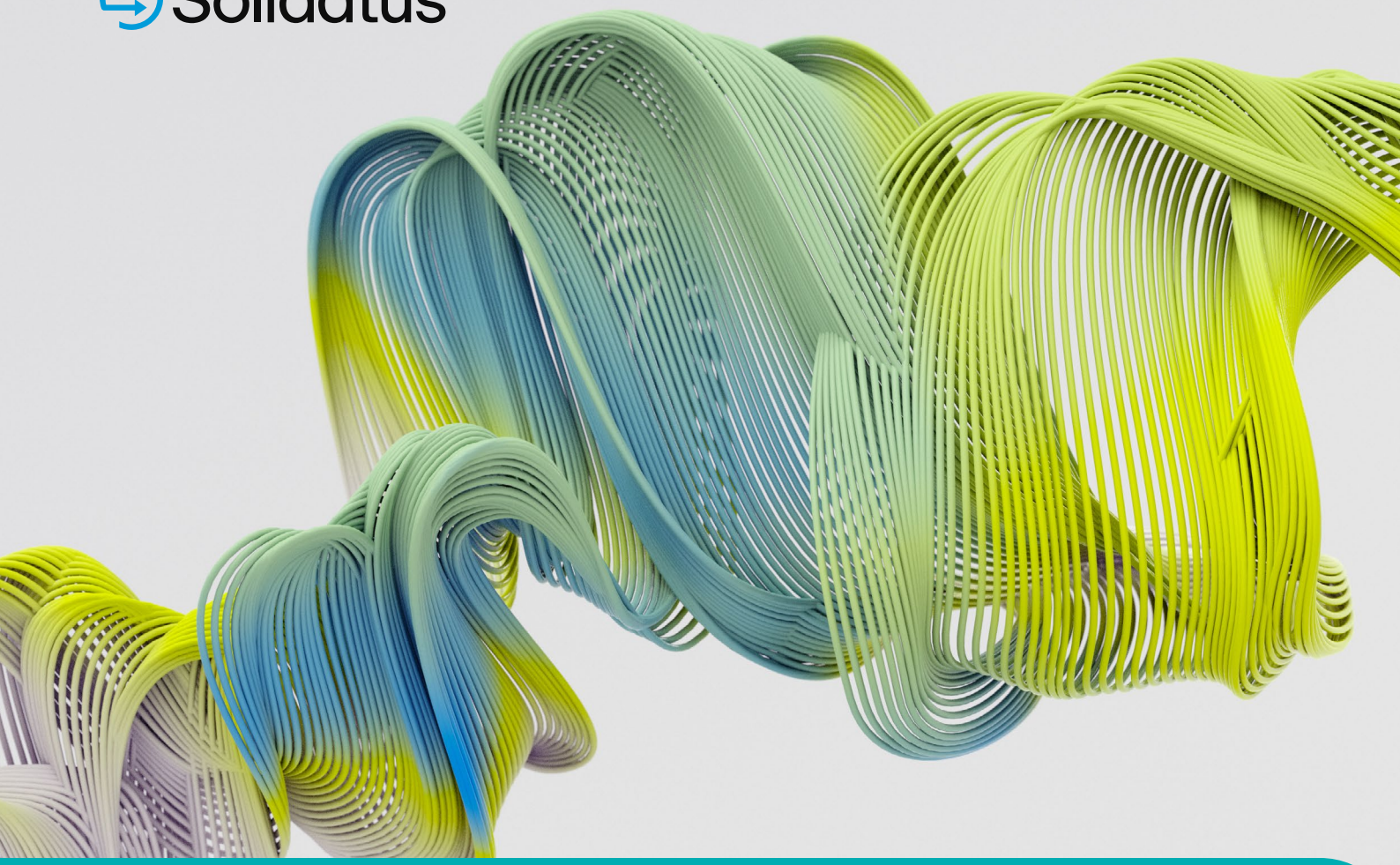
Solidatus



Key concepts in data governance and management

An A to not-quite-Z guide

Contents*

*Please use hyperlinks to navigate to chapters.

Introduction

'Oh, an A to Z guide of key concepts in data governance and management?' we hear you say sceptically. 'Bit ambitious, isn't it?'

Well yes, it would be. But this is an A to *not-quite*-Z guide.

We'll be honest: the title is really just a quirky hook to draw people in our space into a collection of short discussions on some of the terms that most interest us and have captured our attention in recent years.

Some are new – think data mesh and active metadata; some are as old as computing – think data. All are good to acquaint or refamiliarize yourselves with, or to simply compare your take on these terms with ours.

What this isn't

This isn't an anywhere near exhaustive list of every concept in data governance. We barely scratch the surface. To come close to finding that, we'd suggest you start on Wikipedia's data governance page and see where it leads you.

What this is – and what we're including

At Solidatus, we operate in a dynamic space in which our data lineage and catalog capabilities cater for a variety of business, data and tech professionals engaged in numerous data-related activities.

We find it useful to group these activities across:

- Governance and regulatory compliance;
- Data risk and controls;
- Business integration;
- Data sharing; and
- ESG (environmental, social and governance).

(There's a bit more detail on these at the end of this document.)

So we naturally gravitate towards including terms that crop up in these areas. But we expect you'll find them interesting and useful too. We also include one or two terms because they'd be conspicuous by their absence.

A note on how entries in this guide are arranged

We've arranged terms alphabetically but, except for where it appears as its own entry, we ignore the word 'data' in determining an entry's position. So '*graph* database' follows 'data *governance*', for example, but precedes '*data lake*'. You get the idea, so let's dive in.

Active metadata

The way we see it, active metadata is the facility to draw inferences from, visualize dynamically and gain continuous insight from information about data, data systems, business entities and business concepts, the relations between them, and stored knowledge about them.

So, what makes active metadata active? At Solidatus, we'd answer these questions with four points:

- Active metadata includes logical reasoning.
- Active metadata offers a very dynamic form of visualization.
- The information in active metadata is not just about the entities themselves, but about the connections between them; and
- Active metadata should include stored knowledge. This is subtly different from other metadata, because it sits at a higher level, and offers more general – or more universal – information, such as business definitions.

The consequence of all of these is continuous insight. It's more dynamic, it's more complete, it's based on context as well as content, and it respects standards. It's a whole different ballgame.

Let's contextualize active metadata, at least as we see it, in terms of what's delivered, the attributes of an active metadata solution, and – crucially – the main areas for which it can be used.

An active metadata solution:

- Is embedded within an organization's data and business practices.
- Presents a continuous, coordinated, enterprise-wide capability; and
- Provides monitoring, insight, alerts, recommendations and design.

An active metadata analytics workflow:

- Is integrated, managed and collaborative; and
- Orchestrates inter-platform metadata assets and cross-platform data asset management.

Active metadata assets are used to create insight solutions which, among other things, enhance: data integration; resource management; data quality management; data governance; corporate governance; regulatory control; risk management; digital transformation; and ESG.

Above all, the benefits of good metadata capabilities boil down to: making business information complete, coherent, informed and logical; delivering faster, richer and deeper insight; keeping everything up to date; and making your processes reliable and responsive.

Data architecture

Data architecture describes the models and artifacts that should connect a business strategy and data strategy with its technical execution. Primarily, data architecture provides a foundation for people and systems to work with data most efficiently.

Historically, companies adopted computer use in a piecemeal manner, usually first in financial areas and then increasingly across the business as the technology, and the applications, developed and improved. Data architecture as a discipline was a reaction to the obvious inefficiencies (disparate formats, multiple competing sources, incomplete visibility of the data estate, to name but a few) caused by the haphazard adoption of computerisation across the enterprise.

Management consultants McKinsey define data architecture as 'Using data effectively and built on a foundation of business requirements.' This is succinct and, crucially, business-centric, but omits (or possibly elides) efficiency, which is a fundamental driver of data architecture, in order to minimise the costs of data management.

In contrast to the 'why' (see **data governance**), data architecture principally considers the 'where', covering areas such as:

- Modeling and design
- Document and content management
- Storage and operations
- Integration and interoperability; and
- Business intelligence.

Note that all of the above assumes a designed rather than implicit architecture. You might be in a position in which your *data management* could be improved, in which case, mapping your systems with *lineage* software will help.

Business glossary

A business glossary lists business terms along with their definitions. Independent of specific vendors or databases, it simply defines an organization or industry's business concepts.

Data glossaries stand in contrast to data dictionaries, which contain descriptions of datasets or data models. But they complement each other and are often used in parallel.

Both data glossaries and data dictionaries will typically be listed and described in **data catalogs**, alongside all the other data assets that are catalogued by the latter.

Data catalog

A data catalog is a comprehensive directory listing that uses metadata to describe all available data assets a business uses across its entire ecosystem, along with supporting information.

It generally includes references to data dictionaries, **business glossaries** and asset inventories, and it can be broken down into multiple catalogs by division or function.

This metadata helps users understand their organization's data and how it can be used. It also enables data governance and management processes.

In the world of Solidatus, a data catalog is a type of reference model (although in many ways, we go beyond just *data* catalogs; our reference models provide the capability to model control frameworks, policies and regulatory requirements).

Building data catalogs is a useful but hugely laborious task. Our conviction, born of painful experience, is that by taking a **lineage**-first or lineage-enabled approach, you can quickly identify which areas of your data assets you should prioritize cataloguing. Just starting at the top and working your way down is hugely time-consuming and inefficient.

Solidatus lineage models and reference models 'talk' to each other and can be interlinked, making this process that bit easier.

Also see **data glossary**.

Data

OK, let's be honest, everyone knows what data is, if not necessarily in an electronic sense. If you're reading this, it's likely to be your field. But an A to Z guide on data management without an entry on data itself might raise a few eyebrows. So here's our take.

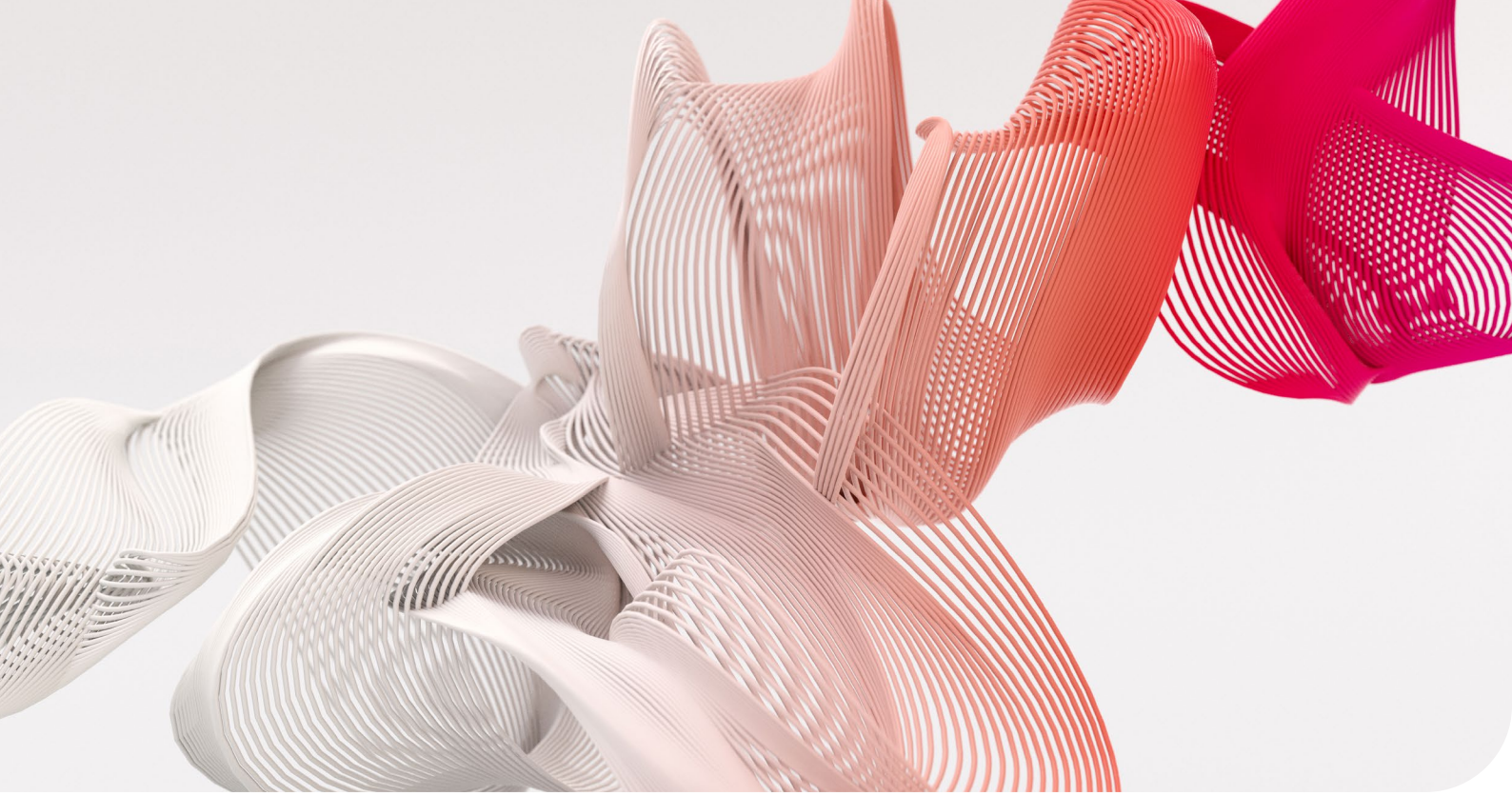
In the broadest sense, data points are facts that are held for a purpose and used as a basis for reasoning and decision-making. In many cases, they arise from business activities; the 'books and records' of a company are now almost universally held as data in computer systems.

Although all data held by an organization should be held for a purpose (you'd hope they wouldn't have been captured otherwise!), some have an essential role in those books and records, either as part of their accounts or as a constituent of a regulatory report. These are referred to as 'critical data elements' (CDEs) to ensure that everybody involved in their generation is aware of their importance to the enterprise.

Databases are aggregations of related data: they may be related by their nature (a database of pedigree dog breeds, for example) or by their purpose (a database of a company's customers). At a technical level, the most common type is the relational database. Often designed to accommodate SQL searches for querying and maintenance, these are traditional and structured, in contrast to **graph databases**, a more recent invention, to which we devote an entry.

We'll save longer discussions on terms – like 'master data', your single source of common data across multiple systems and processes; and 'Big Data', which refers to massive and complex datasets that need special software to deal with – for future pieces of content.

Of course, data differs from *information*. But where do you draw the line?



We've all seen that Lego brick meme with a hierarchy that moves through data, sorted data, arranged data, data presented visually, and data explained with a story; and the higher up that tree the data is, the more likely it is you can describe it as information.

But some data that's right at the bottom still falls into that category, because in the world of data management as opposed to computer science, we'd probably consider something like 'John' to be the smallest unit of data rather than one of the letters in his name, and his name is meaningful.

Databases

See **data**.

Data discovery

Data discovery as a phase describes processes in understanding datasets on hand for data integration and/or data analysis. This step occurs in design and should combine technical search from tools with subject matter expertise from people. During data discovery, a high-level view is taken in assessing data preparation or data quality needs.

Data fabric

A data fabric is an architecture design presented as an integration and orchestration layer built on top of multiple, disjointed data sources like relational databases, data warehouses, data lakes, data marts, IoT, legacy systems, etc., to provide a unified view of all enterprise data.

An emerging data management design to provide flexible, reusable and augmented data management, data fabric achieves this with metadata, which drives the fabric design. Important new elements of a data fabric include active metadata and semantic inference.

As with data mesh, which can complement or compete, data fabric is as much methodology as technology, and it can be designed and deployed manually or automatically.

Data governance

Data governance is our bread and butter, and our solution dovetails with many aspects of it.

As a discipline, it involves organizing, managing, and monitoring the integrity and security of data in an enterprise's system. It requires businesses to establish policies and frameworks to facilitate these processes, ensuring that any new or existing data complies with current internal and external regulatory standards such as GDPR.

Put simply, data governance is the discipline dedicated to ensuring that all of your data and data systems are – and can be demonstrated to be – legal, decent, honest and truthful, and that they fit with your business needs.

At its most basic level, data governance is about bringing data under control and keeping it secure.

Successful data governance requires knowing where data is located, how it originated, who has access to it, and what it contains.

Intelligent data governance begins with clearly defined data policies, standards, and processes relating to information security and privacy, data retention, records management, and more.

Whereas **data architecture** deals more with the 'where', data governance (and data management) deal with the 'why' and 'how', with considerations including:

- Data standards and monitoring: data policy, standards, processes, and compliance tracking and metrics.
- Pure data management and governance: data catalog, metadata management, data lineage, master data, control frameworks, data quality; and
- Information protection, classification and data lifecycle: data classification, data retention and lifecycle, data masking, encryption, and tokenization.

But clearly many of these areas overlap because, well, data is very complicated.

Graph database

When we think of graphs, we tend to picture two-dimensional charts with *x*- and *y*-axes, the type we were first introduced to as schoolchildren.

This can be quite misleading when considering graph *databases*, as the 'graph' component of their name refers to a more advanced – and effectively unrelated – area of mathematics.

Graphs in this context are mathematical structures used to model pairwise relations between objects. These graphs consist of nodes (also called vertices or points) connected by edges (also called links or lines).

In a graph database, data is represented as a series of these nodes and edges, as well as properties.

The graph is a dataset or collection of nodes and edges, the edges representing the relationships between the nodes. The relationships allow data in the store to be linked together directly and, in many cases, retrieved with one operation. Though there is no universal query language, querying relationships is fast because they are perpetually stored in the database (unlike a relational database). Relationships can be intuitively visualized using graph databases, making them useful for interconnected data.

There are two common types of graph database:

- **Property graph.** This is a type of graph model where relationships not only are connections but also carry a name (type) and some properties. A property graph excels at showing connections among data scattered across diverse data architectures and data schemas. (The Solidatus solution isn't a graph database *per se*, but it *does* exhibit some associated characteristics, particularly of property graphs. So it can be considered a graph application in that it can read and interact with property graphs.)
- **Knowledge or RDF (resource description framework) graph**. In a knowledge graph, information is represented by separate nodes.

Data lake

A data lake is an environment or storage platform in which a vast amount of data can be ingested, stored, assessed, and analyzed.

It contains all an organization's data in a raw, unstructured form, and can store the data indefinitely – for immediate or future use. A data warehouse contains structured data of various types that has been cleaned and processed, ready for strategic analysis based on predefined business needs

Data lakes serve many purposes, including offering an environment for data scientists to mine and analyze data. They also provide a central storage area for raw data, often with minimal, if any, transformation.

Loaded and stored in a 'raw' format, data in a data lake needs no indexing or preparation, bringing the flexibility to do all manner of analytics on huge and comprehensive datasets in a central location.

Rather than replace them, data lakes can enhance the utility of **data warehouses**.

Data lineage

The concept of **data** lineage is straightforward in theory: it's simply a description – or trace – of the journey of a specific piece of data (or, technically, data element value), one that outlines where it first came into your organization's data ecosystem, where it leaves, its stops along the way and the impact it makes on other data.

But it gets complicated for several reasons, including:

- Changes it undergoes along the way.
- Split paths that a piece of data takes – linear paths are rarely smooth and one-dimensional; and
- The 'temporal' nature of lineage, which calls for version control.

And this is before you look at its many possible downstream journeys and possible duplications, some pre-determined, others up for debate – systems are rarely set in stone.

Let's consider an analogy, one that demonstrates the importance of taking a 'lineage-first' approach:

Picture a lake and the streams that feed into it: the lake is our data; the streams represent the sources of our data.

In a simple world, where you have data in silos or small disconnected 'pools of water' that don't affect anything downstream, then correcting the data quality problems upstream will solve your issues. But almost no organizations today are structured that simply – all contain numerous connected systems, and if a stream becomes polluted, it then carries that pollution into the lake.

We can try and remove the pollution from the lake each time we see 'spills'. But, if we don't identify the source of the pollution and remedy it, our lake will remain tainted.

If we consider the entire water system: currents of water pass through a multitude of streams that feed into the main body of water. And, that water is often connected to a matrix of other bodies of water with their own network of streams.

When we adjust problematic data at its source, how does it affect systems beyond our single data 'lake'? Is it used for regulatory reporting? Is it used for P&L? Is it customer data?

Depending on the size of the organisation, data problems can create a domino effect across countless systems in multiple jurisdictions. And If we can't trace the lifecycle of data from its point of origin, we are stumbling through the dark when we attempt to solve problems.

Lineage provides the root-cause analysis. It allows us to see the entire forest and not just the trees. It illuminates our problems and it allows us to fix the things which have a potentially negative impact on the organization.

The larger the organization, and the more complex and diverse its systems, all of which interact with each other in complex ways, the greater the data headache and the likelihood that is held by local experts rather than shared by the organization – which the right approach to lineage can alleviate.

The best way to make sense of it all is to model your lineage with a software solution that allows you to visualize it and, by applying rules that you set, isolate specific paths that cut out the noise.

Put simply, a lineage model – such as those you can build in Solidatus – is a visualization that shows you the interplay and flow of data (or other types of 'objects' that you map) and the uses to which it is put within a context framed by the modeler.

There are numerous applications for lineage analysis, most of them dovetailing with other entries in this document, including: auditing and compliance; data quality; data governance; and data integration and cloud migration. And we believe that one of the most fundamental, tying in with all the rest, is in smarter **data catalog** work.

Data management

This whole document concerns terms that loosely fall within the broad discipline of data – and metadata – management, so we needn't dwell on its definition.

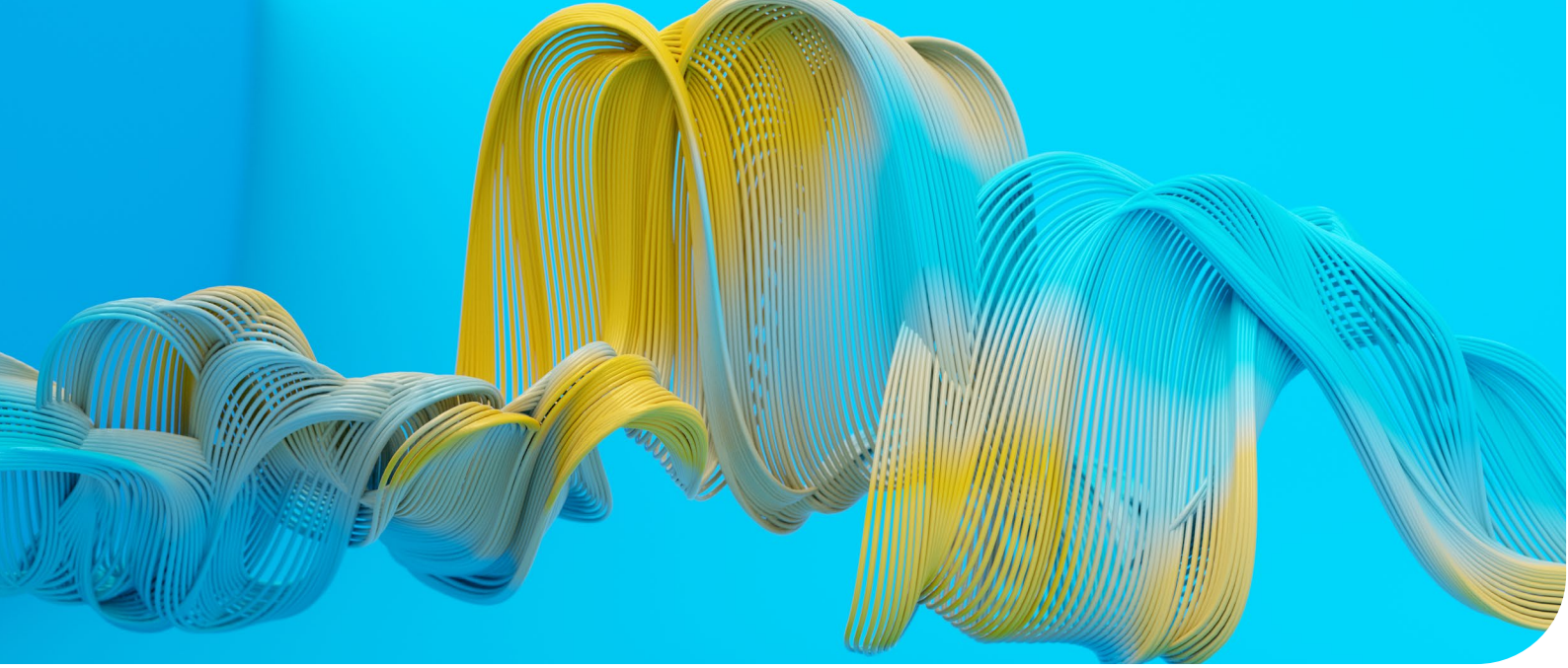
Here's something short and sweet that should cement the context on which this compilation is based:

Data management is the process of collecting, keeping, and using data in a cost-effective, secure, and efficient manner. Data management helps people, organizations, and connected things optimize data usage to make better-informed decisions that yield maximum benefit.

In our world, there's a significant crossover with **data governance**, on which we have a longer entry.

Data mart

A data mart is a subset of a **data warehouse** and is designed to service a specific business line or purpose. Data warehousing pioneer Ralph Kimball conceived of data marts to 'begin with the most important business aspects or departments.'



Data mesh

'Data mesh' is an evocative label. It might sound quite high-tech and abstract. But it's actually quite down-to-earth.

At its core, a data mesh is simply a set of principles for designing a modern distributed **data architecture** that focuses on business domains – not on the technology used – and treats data as a product.

As an architectural pattern for data management and access, it emphasizes decentralized ownership, standardization, and collaboration. In a data mesh, data is treated as a first-class citizen and is made available to all parts of the organization through a set of standardized, domain-specific data products.

This approach enables teams to independently develop and own their data products, while still ensuring that data is consistently managed and governed across the organization. A data mesh also typically includes mechanisms for data discovery, collaboration and governance, and it can support a wide range of data-driven applications and use cases.

The goal of a data mesh is to provide a scalable, flexible, and self-serve data infrastructure that can support the evolving needs of an organization, making them less reliant on centralized resources like **data lakes**.

A data mesh approach introduces multidimensional technical, cultural, organizational and other paradigm shifts for companies that formerly used a centralized data management architecture.

So, though there are products that support it, it's as much a methodology as a tangible thing. It's about data discoverability: the concept, the principles and how to do it.

This methodology is founded on four key concepts:

- Data ownership that is domain-centric.
- Treating data as a product.
- Facilitating instant – self-service – access to data; and
- Distributed **data governance**.

Clearly, the better your software solution, the easier it is to navigate, map and discover components of your data mesh.

You'll often hear data mesh discussed alongside **data fabric**, the suggestion being that organizations should adopt an either/or approach. But the two paradigms can happily coexist in some circumstances.

Metadata

Metadata is arguably more 'interesting' than data. It can tell us more than data itself. With the right tools, it's limitless in its scope. And it's the lifeblood of solutions such as Solidatus. But what is it beyond the tired, cringe-inducing cliché of being 'data about data'?

Well, we'd define this as a special kind of data that describes business processes, people, data and technology, and the relations between them, bringing context and clarity to the decisions that link them. Traditional examples include the metadata in a **data catalog** or **business glossary**.

Put another way, metadata is data that's been recorded in IT tools and which improves both the business and technical understanding of the data and data-related assets.

In essence, metadata is data *documentation*.

Still want more? Well, if you skipped over it at the start of this guide, you should note that we really go to town in our entry on **active metadata**, the metadata that everyone should be talking about.

Data model

'Data model' can mean several things. It can be used as a noun and a verb. And it has applications well beyond data management and data governance. But in almost all cases, it's something that explicitly determines the structure of data and how elements relate to each other. Or, in less ideal worlds, it describes data and systems after the fact, if they've already been structured without a plan before you model them.

At Solidatus, one way we think of data models is as blueprints. So what is a blueprint?

In the sphere of construction, a blueprint is a two-dimensional drawing that provides a visual representation of a building's layout, dimensions, component placement, electrical wiring and construction materials. Drawn up by architects or engineers, blueprints allow you to quickly check and identify different building elements and verify compliance with building codes.

A blueprint follows a carefully thought-out, logical sequence of steps. Drawing pages in blueprint sets are arranged in a predictable fashion, and blueprint symbols and lines have highly specific meanings.

Imagine if instead of a building inspection, you were required to undergo a business inspection.

Could you visually lay out the dimensions and components that inform your business decisions and drive your operations?

Are you confident that you have structural integrity, and that your data governance and management capabilities provide the appropriate level of safety and soundness required to verify compliance with regulatory obligations, reporting requirements and privacy standards?

Traditional **lineage** approaches can be daunting for any organization. Having to source and then document the sources of information that describe the processes, people, systems, data, reports and controls seems unattainable. Especially considering that everything that needs to be documented has likely never been documented.

With the right lineage software solution, you can ingest metadata across key systems, providing a view of the federated collection and documentation of key processes that drive your business.

Put simply, a blueprint in the context of data management is an interactive, living visualization of how your data flows and its connection to the obligations that regulate it, your policies that guide it and your processes that create or use it – both now and at other points in time.

We focus here on lineage models but this thinking and applies to reference models, which, in Solidatus terminology, can be used as a type of **data catalog**.

It's worth considering these ideas alongside data supply chains. These represent the technological steps and human-involved processes supporting the flow of data through the organization, from its raw state, through transformation and integration, all the way through to the point of consumption or analysis. By this definition, the data supply chain comprises three distinct data phases: raw data; transformation and integration; and consumption and analysis.

Ontology

An ontology is a set of concepts and categories in a subject area or domain that shows their properties and the relations between them and a way of showing the properties of a subject area and how they are related, by defining a set of concepts and categories that represent the subject.

Within data governance, this creates a common understanding of the data domain and helps create a shared understanding of the scope and internal relationships.

The main components of an ontology would usually be described as:

- **Classes:** the distinct types of things that are in our data or datasets;
- **Relationships:** the properties that connect two classes together; and
- **Attributes:** the properties that describe individual classes.

With an ontology, you're in effect taking a **taxonomy** to a broader level in that, while you're dealing with only one conceptualization, it can encompass multiple taxonomies.

Taxonomy

In its broadest sense, a taxonomy is a scheme of classification, especially a hierarchical classification, in which things are organized into groups or types.

This applies equally in the world of data.

In this world, having a shared taxonomy or set of agreed classifications – or categories – is vital in data governance to ensure that all users are discussing the same data in the same context. Typical examples would include those in the EDM Council's DCAM model (see **bit.ly/aboutdcamreview**) or in the EU's ESG taxonomy (e.g. **bit.ly/taxonomymapping**).

See the related concept of 'ontology' above.

Data warehouse

A data warehouse is a repository used to provide decision-support data and aid workers engaged in reporting, querying, and analysis. This architectural technology enables organizations to integrate data from a range of sources into common data models.

Arising to address the high costs of extracting and analyzing data from organizations' operational systems, data warehouses store data that's both highly curated and easily available to be used by data analysts, business analysts and developers.

Their utility can be enhanced by **data lakes**.

Zeroes and ones

Zeroes and ones are the smallest units of digital data, representing 'off' and 'on', and used in machine code. While we're all entirely reliant on them, they fall well outside of the usual day-to-day life of data practitioners. But an A-Z guide without an entry for Z would raise a few eyebrows, so we're throwing this in for free.

About Solidatus and our solutions

Solidatus is an innovative data management solution that empowers organizations to connect and visualize their data relationships, simplifying how they identify, access, and understand them. With a sustainable data foundation in place, data-rich enterprises can meet regulatory requirements, drive digital transformation, capture business insights, and make better, less risky and more informed data-driven decisions.

We provide solutions to several key areas of endeavour, including: governance and regulatory compliance, data risk and controls, business integration, environment, social, governance (ESG), and data sharing. Our clients and investors include top-tier global financial services brands such as Citi and HSBC, healthcare, and retail organizations as well as government institutions.

solidatus.com

More on our solutions

Cast your mind back to our introduction and you'll remember that we mentioned our five 'solutions'.

Of course, Solidatus is just one piece of software, so these solutions – or use cases – simply provide different lenses through which to view them. But they provide a useful way of assessing the various objectives of people working with data. What does this mean?

Well, we provide a low-touch solution to complex data problems, with our 'lineage-first' approach, proven to offer up to 90% efficiency savings over traditional methods.

Solidatus allows businesses to elevate their organizational data culture and capabilities by enabling the creation of a holistic organization-wide digital map that details all the relationships that interact with and have an impact on their data, accelerating modernization and transformation.

Solidatus is the first choice for data-rich organizations and is being chosen for a growing number of vital client initiatives including digital transformation, data governance, cloud migration and regulatory compliance.

So let's finish with an overview of our specific solutions:

Data governance and regulatory compliance

Data is your business's most critical asset, but it carries with it significant risks. Data that is not secure, accurate, documented, managed, and audited leaves you vulnerable – at risk of costly regulatory fines.

Using Solidatus, you create living blueprints that map how your data flows as it moves through your systems – both now and at other points in time. We connect your data to the processes that create it, to the policies that guide it, and to the obligations that regulate it. With this framework in place, you can maintain transparency across your business, meet ever-evolving regulatory requirements, and accelerate change programs.

Data risk and controls

Solidatus extends beyond traditional data governance software by adding key business dimensions including glossaries, definitions, ownership, privacy and authority, which are all visualized in context and maintained in Solidatus through both business and technology lenses.

With dynamic connections formed between your data flows and information risk and control frameworks, you can then embed processes that identify data-related risks and enforce accountability for the implementation and monitoring of data controls.

Business integration

Your enterprise is becoming more and more complex. Disparate, disconnected infrastructures and data siloes end your business transformation projects before they begin, stopping you from capturing unique insights and new revenue streams.

Solidatus unifies the pillars of your business and shines a light on your data fabric, allowing you to create interactive and living blueprints that visualize how your data flows, and its connections to the processes that create it, to the policies that guide it, and to the people that use it – both now and at other points in time. From this vantage point, an unlimited number of questions can be answered, and the outcomes analyzed.

Data sharing

All organisations need to share data to deliver effective services to clients, but this carries significant financial and reputation risk if you can't demonstrate this sharing is controlled, understood and compliant with legal requirements and client permissions. These data sovereignty-and data residency-related risks are exacerbated when organisations span jurisdictions.

To demonstrate control of their data, organisations must show they understand each data sharing process, including those relating to: client location, destination of data transfer, type of data (including whether it's personal or sensitive personal data), the purpose of sharing it, client and data owner permissions, jurisdictionally varying legal requirements, and internal agreements. You need a solution to make sense of this spaghetti of interrelated issues across jurisdictions.

ESG (environmental, social and governance)

Solidatus brings ESG priorities, reporting and assessments together with the data flows, processes and their owners in a managed, sharable and maintained environment. Better governance, less waste, more efficiency and more effective socialization of ESG programs make Solidatus the best solution for meeting ESG regulations.