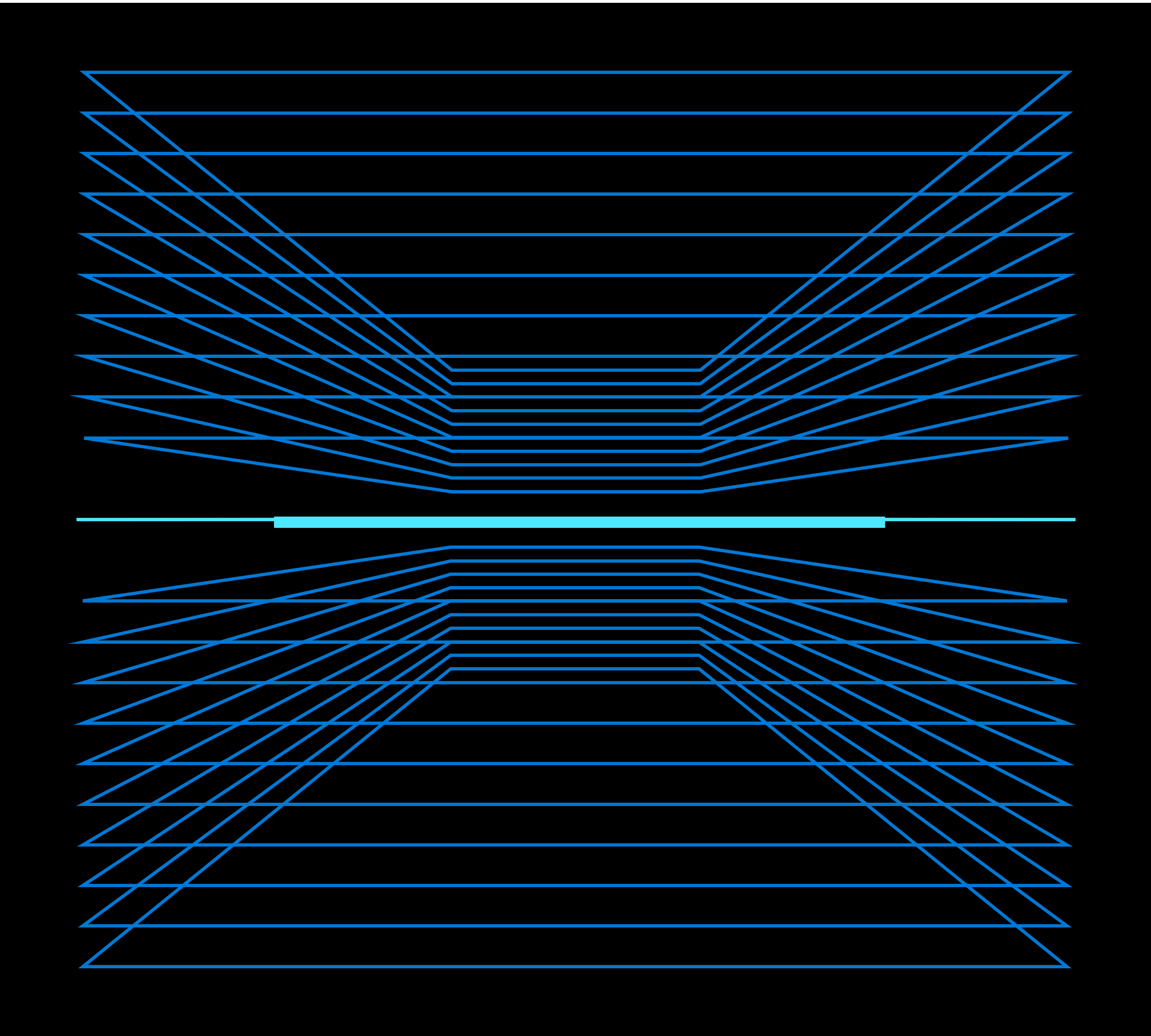
Microsoft

A Guide to Data Governance

Building a roadmap for trusted data



Contents

What is Data Governance?

Data Governance is about ensuring that the data being used in your core business operations, reports and analyses is protected and can be trusted

In many companies today, data governance has become increasingly important but what exactly is it? What does data governance mean?

There are several definitions of data governance available from various sources. These include the following quotations:

> "Data Governance (DG) is defined as the exercise of authority and control (planning, monitoring, and enforcement) over the management of data assets."
>
> Source: DAMA Data Management Body of Knowledge V2 (DMBOK2)

> "Data governance encompasses the people, processes, and technology required to create a consistent and proper handling of an organization's data across the business enterprise."
>
> Source: Wikipedia

> "Data governance is the orchestration of people, processes, policies and technology to formally define, discover, assess, clean, integrate, and protect structured and unstructured data assets through their lifecycle to guarantee commonly understood, trusted and secure data throughout the enterprise"
>
> Source: Mike Ferguson, Intelligent Business Strategies.

Looking at these, data governance is about ensuring that the data being used in your core business operations, reports and analyses is discoverable, accurately defined, and is totally trusted and can be protected. Additionally, data governance has become increasingly important to a business because according to a prediction from IDC, digital data is expected to grow to approximately 175 zettabytes by 2025.[1]

But why is data governance so important? Why is it needed?

[1] https://www.idc.com/getdoc.jsp?containerId=AP45214519

Why Do We Need It?

There are many reasons why data governance is needed. These include the need to govern data to maintain its quality as well as the need to protect it. This entails the prerequisite need to discover data in your organization with cataloguing, scanning, and classifying your data to support this protection.

The Need to Create Trusted Data

Data governance is needed to improve data quality so that data is trusted.

Data quality if of utmost importance because when companies work with inferior data, this negatively impacts their downstream insights, analyses, and recommendations. Data quality must entail the data is complete, unique, valid, timely, accurate, and consistent.

Data quality problems can impact on business operations causing process errors, process delays, unplanned operational costs and inaccurate decisions.

Data needs to be governed across a distributed computing environment

In many companies today, the expectation in the board room is that data and artificial intelligence (AI) will drive competitive advantage. Not surprisingly therefore, executives are eager to sponsor AI initiatives in their determination to become data driven. However, for AI to become effective, the data it is using must be trusted. Otherwise decision accuracy may be compromised, decisions may be delayed, or actions missed which impacts on the bottom line. Companies do not want 'garbage in, garbage out'. It might seem relatively straight forward to fix data quality until you look at the impact that digital transformation has had on data in the last few years.

For most companies, the introduction of digital transformation has resulted in a more complex operating environment in comparison to just having a single data centre. Today, most companies have created an operating environment that spans the edge, multiple clouds and the data centre. Surveys over the last couple of years have shown this with one[2] last year showing 81% of companies surveyed had systems running in multiple public clouds and one or more private / dedicated clouds. That typically translates to meaning that both operational and analytical systems are running in the cloud and the data centre. Examples of operational transaction processing systems running in the cloud include Microsoft Dynamics, Workday, Salesforce, ServiceNow and Marketo. Analytical systems running in the cloud could include data warehouses, graph databases, data lakes being used by data scientists and real-time IoT streaming analytic applications. The result is that companies are now dealing with a hybrid environment with data in multiple different data stores that are scattered across all of this landscape similar to that shown in Figure 1.

[1] https://blog.syncsort.com/2019/01/data-quality/data-integrity-vs-data-quality-different/

[2] IDC's Multi-cloud Management Survey 2019

The distributed data estate

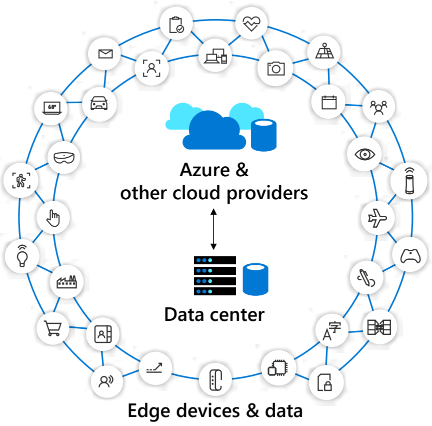


Figure 1

This includes data stored in edge databases relational DBMSs, NoSQL DBMSs, Files, Cloud storage, Hadoop systems and scalable messaging queuing systems (e.g. Kafka).

Digital transformation has resulted in a lot of new data sources that businesses want to analyse

The other major impact of digital transformation is that there are a lot of new data sources that business now wants to analyse beyond the traditional master data and transaction data found in data warehouses. This includes machine generated data such as clickstream data in web server log files, human generated data from social networks, inbound email, and open government data. Also, unstructured content in various documents is in multiple locations.

Data is increasingly spreading out across data stores in the data centre, multiple clouds and at the edge which makes it harder to find and harder to govern

With data increasingly spreading out across a hybrid multi-cloud, distributed data landscape, it is not surprising that people are struggling to know where it is in order to govern it. Yet, the business impact from ungoverned data can be considerable. Poor data quality impacts business operations because data errors cause process errors and delays. Poor quality data also impacts business decision making and the ability to remain compliant. Data governance needs to therefore include data discovery, data quality, policy creation, data sharing, and metadata to help track and govern data activity.

The Need to Protect Data

Data needs to be protected to prevent data breaches and enables you to remain compliant on data privacy with regulations and legislation

The other major driver for data governance is data protection. This is needed primarily to remain compliant on data privacy with regulatory legislation such as the European Union General Data Protection Regulation (GDPR) or the California Consumer Privacy Act (CCPA) and to prevent data breaches. Data privacy as well as the growing number of data breaches has made data protection a top priority in the boardroom. These breaches highlight the risk to sensitive data such as personally identifiable customer data. The consequences of data privacy violation and / or a data security breach are numerous and include:

- Loss or serious damage to brand image
- Loss of customer confidence and market share
- Fall in share price which impacts stakeholder return on investment and executive salary
- Major financial penalties as a result of audit/compliance failure
- Legal action
- The ‘Domino effect’ of the breach, e.g. Customers may also fall victim to identity theft as a result of a breach

Also, in most cases, publicly quoted companies must declare these breaches.

If they happen, customers are more likely to blame the company first rather than the hacker and may boycott the company for several months or may never return.

Failure to comply with regulatory legislation like GDPR and CCPA on data privacy may also result in several very significant financial penalties. No one wants any of this and so governing data to avoid such risks is well worth doing.

Given this backdrop, what then are the requirements to govern data in a modern enterprise?

Requirements For Governing Data In A Modern Enterprise

The requirements for governing data include:

- Data item and data entity definition to create a common business vocabulary in a business glossary
- Data item and data entity identification / discovery
- Data governance classification to govern data access security, data privacy and data retention
- People - including data owners with governance accountability and data stewards responsible for protecting it and upholding its quality
- Data governance processes
- Policies and rules to define how specific data should be governed throughout its lifecycle
- Policy enforcement across data stores in the distributed data landscape
- Master data management to get data consistent across operational and analytical systems e.g. customer, product, supplier
- Metadata lineage
- Technology to make it possible to govern structured, multi-structured and unstructured data across data centre, multiple clouds and the edge

There are a number of requirments that need to be met to successfully govern data

Common Business Vocabulary

A common business vocabulary is a set of commonly defined data names and data definitions documented in a business glossary within a data catalog

It's purpose is to ensure that data is consistently named and commonly understood especially when it is shared

The first of these is critical. It is a common business vocabulary. The reason this is needed is to clear up ambiguity in the meaning of data caused by ambiguous data names in different data stores, in reports and in data made available through APIs. For example, a business analyst may produce a report that shows "Total Sales". Another report from a different analyst may show "Sales" and another uses "Gross Revenue". Are they the same? Are they different? Does Total Sales include sales tax or not? What about Sales? Does that include sales tax or not? These are everyday ambiguities that need to be avoided. The answer lies in creating a common business vocabulary of common data names and data definitions for data entities and their attributes clearly defined in a business glossary.

Governing Data Across A Distributed Data Landscape

Data quality, data privacy, data access security and data retention need to be governed across the data centre, multiple clouds and the edge

However, a lot more is required. Enterprise data governance is needed across data created and stored on-premises, in multiple clouds and at the edge. That means being able to govern data quality, data privacy, data access security and data retention across this landscape and also provide full metadata lineage. It also means that other questions about data need to be answered. For example:

- What data exists across this landscape and where is it stored?
- What data needs to be governed and managed?

- How should it be classified? For example, is it sensitive data such as personally identifiable information (PII), is it a trade secret?
- If the data is structured, what data names is it known by and are there any common data names that it should be known by? Also, is the same data stored in different data stores with different data names?
- How good or bad is the quality of the data?
- Does it need to be cleaned, transformed, integrated and shared?
- Who is responsible for doing that work?
- Is there an owner for any of this data?
- What trusted data is available and how was it produced?
- If the data changes should it be kept synchronised?
- If the data is master data, then who is allowed to access and change?
- Who creates policies to govern specific data?
- Do changes to these policies need to be approved and if so by who?
- How much power do users have and how are users, applications and scripts audited?

Many questions about data need to be answered in order to govern it

The other challenge is that data is being collected and stored in multiple places across the enterprise. This may include data collected and stored in different geographies and different legal jurisdictions. As a result, different legislation may apply to governing the same data in different jurisdictions. You therefore need to discover what data exists across the hybrid multi-cloud distributed data landscape (including geographic location) to be able to:

Data needs to be governed across multiple geographies that may be subjected to different legislation

1. Understand what data attributes, data entities and data relationships exist across the distributed data landscape
2. Classify the data to know how to govern it
3. Define policies to specify how data should be governed for each type of governance classification
4. Enforce data quality, data access security, data privacy and lifecycle management policies across the distributed data landscape

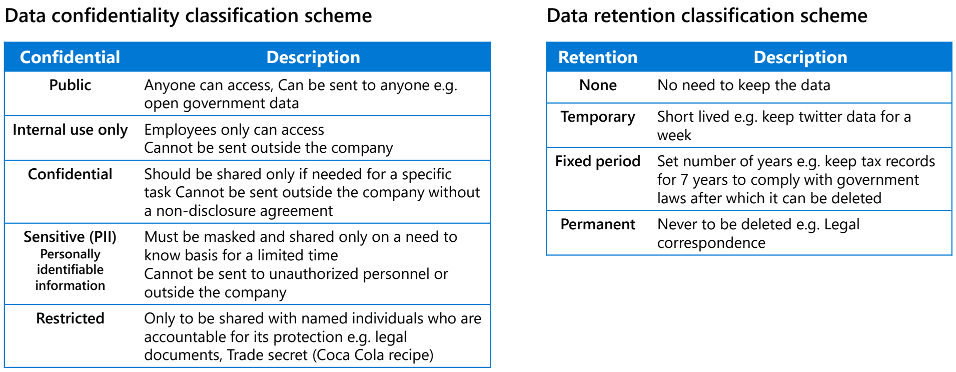
Data Governance Classification

In addition, there needs to be some way to classify data to understand its level of confidentiality and how long to retain it for. This requires:

- A data confidentiality classification scheme
- A data retention classification scheme

An example of each of these schemes is shown in Figure 2:

Data classification schemes are needed to govern confidentiality and retention



Data confidentiality classification scheme

Confidential	Description
Public	Anyone can access, Can be sent to anyone e.g. open government data
Internal use only	Employees only can access Cannot be sent outside the company
Confidential	Should be shared only if needed for a specific task Cannot be sent outside the company without a non-disclosure agreement
Sensitive (PII) Personally identifiable information	Must be masked and shared only on a need to know basis for a limited time Cannot be sent to unauthorized personnel or outside the company
Restricted	Only to be shared with named individuals who are accountable for its protection e.g. legal documents, Trade secret (Coca Cola recipe)

Data retention classification scheme

Retention	Description
None	No need to keep the data
Temporary	Short lived e.g. keep twitter data for a week
Fixed period	Set number of years e.g. keep tax records for 7 years to comply with government laws after which it can be deleted
Permanent	Never to be deleted e.g. Legal correspondence

Figure 2

Automating the data confidentiality and data retention classification process using the classes defined in each scheme is needed to consistently label data across the distributed data landscape to enable it to be consistently and correctly governed. Rules and policies would then need to be defined for each class in the classification scheme to specify how to govern data according to its classification.

Data Governance Roles and Responsibilities

Another requirement is the need for accountability. Without this, confusion lingers as to who is accountable for governing data. If there is no accountability, how do you answer the following questions?

- Who sets success metrics and monitors how well the data governance program is working?
- Who are the data owners?
- Who defines and maintains a business glossary?
- Who creates and maintains policies on access security?
- Who is protecting PII data privacy for compliance with GDPR and CCPA?
- Who is looking after the quality of product data across all brochures and partner websites?
- Who ensures customer data is consistent across all systems?
- Who is policing external subscription data usage Vs the license?
- Who is policing privileged users like DBAs and data scientists?

Is it a C-level executive? Is it a department head? Is it the head of governance, risk and compliance? What about the legal department? Or is it IT's responsibility? Roles and responsibilities are needed to avoid confusion and to set the foundation upon which a data culture can materialize.

Data Governance Processes

Data governance processes are needed to govern how data is defined, discovered and classified

In addition to roles and responsibilities, processes are needed. For example, to:

- Govern the definition and maintenance of a common business vocabulary
- Discover and identify what data you have, what it means and where it is
- Classify data to know how to govern it

Data governance processes are also needed to ensure governance policies and master data are created and maintained correctly

- Govern the definition and maintenance of data access security policies
- Govern the definition and maintenance of data privacy policies
- Detect data quality problems and remediate them
- Apply policies to ensure action is taken for compliance
- Govern maintenance of master data

Data Governance Policies and Rules

We also need to define policies and rules to govern:

Policies and rules are needed to ensure data is protected and kept in the highest quality throughout its lifecycle

- Data integrity
- Data ingestion
- Data access security
- Data privacy
- Data quality
- Data maintenance
- Data retention

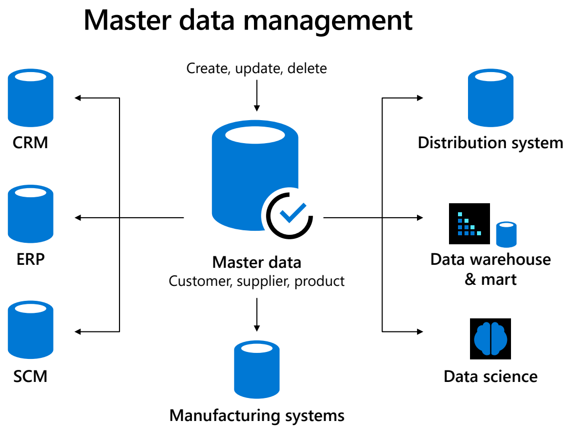
These need to be associated with each class in the aforementioned data governance classification schemes.

Master Data Management

Another central requirement in governing data is master data management (Figure 3). Master data is the most widely shared data in any organization and includes core data entities such as Customer, Supplier, Materials, Employee, Asset. It also includes financial Chart of Accounts data that is found in different financial applications.

Because master data is so widely shared it is application agnostic. It is needed by both operational transaction processing applications and analytical systems. Keeping this data synchronized can resolve so many data errors and process errors. Therefore, maintaining it centrally via a common process and synchronizing every system that needs it is the ideal situation. In addition, governance is needed over who is allowed to maintain it and where that maintenance needs to happen.

Master data management is needed to ensure that master data is centrally maintained and synchronised across all operational and analytical systems



CRM = Customer Relationship Management
SCM = Supplier Chain Management
ERP = Enterprise Resource Planning

Figure 3

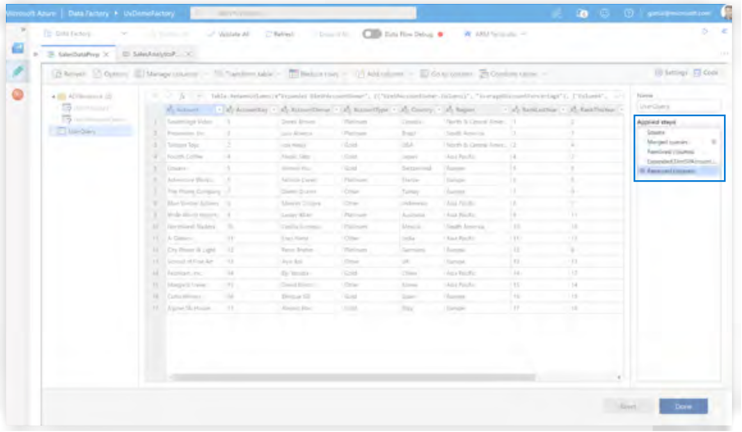
The same applies to reference data such as code sets and financial markets data. In this case standardization and synchronization of code sets is known as reference data management which is also a requirement.

Metadata Lineage

Finally, there is a requirement for metadata lineage (Figure 4). This is the need to provide an audit trail to know where data originated and how it has been transformed on route to a report or a data store. In addition, metadata is needed to trace who or what is maintaining data (e.g. master data) including when and where this occurs.

Metadata lineage provides an audit trail on where data originated, how it was transformed on its way to the point of use and how it has been maintained

Metadata lineage



Tracks all activity performed on the data by the user

Figure 4

What Is Needed For End-to-End Data Governance?

To address these requirements, we need an end-to-end solution that is capable of governing data throughout its lifecycle across data stores in the edge, multiple clouds and the data centre. This is shown in Figure 5.

Data governance framework

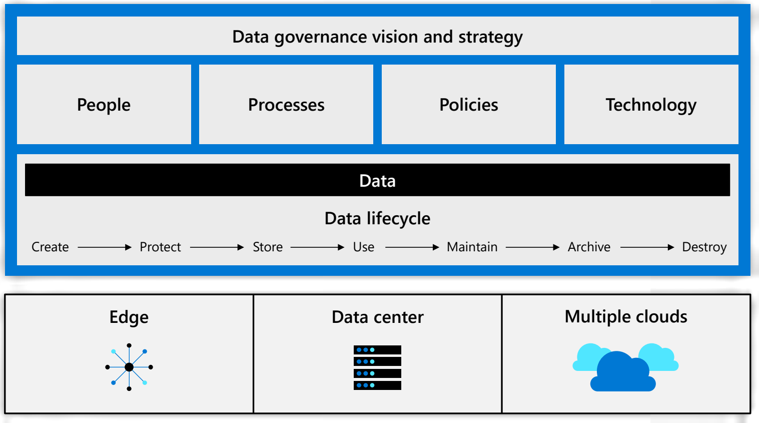


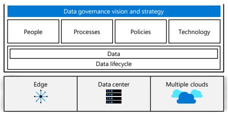
Figure 5

The solution consists of several components:

- A data governance vision and strategy
- The data itself e.g. customer data, supplier data, orders data etc.
- The data lifecycle from creation to destruction within which data needs to be governed
- Data governance roles and responsibilities (people)
- Data governance processes and activities and how they apply to the data lifecycle
- Policies and rules to govern data at different points in the lifecycle
- Data governance technologies to help make this possible

Components Needed For Data Governance

Data Governance Vision and Strategy



A data governance strategy should specify objectives, success metrics and targets to be achieved

At the top of the solution is the data governance vision and strategy. This includes a vision statement, the stakeholders backing the data governance program and the objectives of the program. These objectives should be aligned with strategic business objectives to show contribution to common goals. The strategy also includes success metrics (KPIs) and targets to be reached to monitor the progress. There are two types of metrics to be considered. The first is risk management and compliance metrics designed to measure improvements in data quality, security, privacy and retention. The second type is value creation metrics. These help to monitor how data governance is contributing to improving business value through the creation and use of trusted data. Business value in this case could mean:

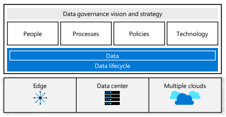
- Reducing risk (e.g. by protecting against a data breach)
- Reducing costs (e.g. by eliminating data errors in business processes that cause unplanned costs to mount as people step in to fix them)
- Increasing revenue (e.g. by providing high quality integrated and trusted data that improves accuracy of next best offer recommendations to drive up revenue).

It should also specify business cases which are often best articulated by describing the impact ungoverned data is having on the business

The data governance strategy should also include business cases which are best articulated by describing the impact that *ungoverned* data is having on the business. Describing the business problems caused by ungoverned data helps to systematically identify candidate business cases. It also allows you to rank the business problems in order of severity and return on investment (ROI) if the problems were solved. Prioritising problems where ungoverned data has the greatest business impact is an extremely effective way to getting stakeholder sponsorship. This is because it pinpoints the greatest opportunities to drive value.

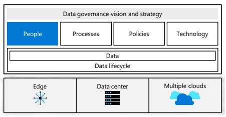
The data governance strategy should then include the projects / initiatives that are needed to achieve the business objectives, meet the targets set and deliver the ROI identified in business cases. In addition, it should include the budget allocated to these projects, who is leading the data governance program and who is accountable for achieving them. It may also include some data principles. Two examples here are that data should be treated as an asset and that data is the property of the company and should be shared.

Data and the Data Lifecycle that Needs to be Governed



With respect to data we mean data entities, documents, unstructured images, video and audio. Examples of data entities are customer, product, employee, supplier, order, invoice, payment and asset. Examples of documents are a supplier contract, an annual report and a product brochure. The data governance solution should enable you to govern data throughout the lifecycle. That means governing data creation / ingestion protection, storage, use, maintenance, archiving and destruction.

Data Governance Roles and Responsibilities Guidance (people)



With respect to people, there are a number of data governance roles and responsibilities. These can vary across organisation and so the follow roles and responsibilities listed in the table below are provided as guidance only.

Role	Responsibility
Executive sponsor (e.g. CFO / CIO)	Senior business stakeholder with authority and budget who is accountable for ensuring data governance is established
Data Governance program leader (e.g. CDO or appointed lead)	The person with overall accountability and responsibility for implementing the data governance program.
Data Governance Control Board	Includes data governance lead and data owners. Sets success metrics, owns the data governance roadmap, selects working groups, holds the budget for the data governance program, arbitrates when conflicts occur on priorities and definitions of cross functional data
Data Governance Working Group	Plan and progress data definition and improvement of a specific data domain (E.g. Customer or Supplier), update Data Governance Control Board on progress, manage stewardship across the enterprise for a specific domain
Data owner	Senior business stakeholder with authority and budget who is **accountable** for overseeing the quality and protection of a specific data subject area or data entity across the enterprise and make decisions on who has the right to access and maintain that data and on how it is used
Business data steward	Business professional **responsible** for overseeing the quality and protection of a data subject area or data entity. They are typically experts in the data domain and work in a team with other data stewards across the enterprise to monitor and make decisions to ensure data quality is maintained
Data Protection Officer (DPO)	Senior business stakeholder with authority and budget who is **accountable** for the protection of personal data specific to compliance legislation in all jurisdictions that the company operates
Data security team	Responsible and accountable for data access security and data privacy policy enforcement
Data Publishing Manager	Responsible and accountable for quality assurance checking and publishing of newly created trusted data assets in a data marketplace for consumers to find and use

Roles, responsibilities and organisational structure set out how to organise people to successfully govern data

The objective is to organise in a way that allows you to take a 'divide and conquer' approach to governing data throughout its lifecycle across a hybrid computing environment. One way of doing this is to have multiple working groups reporting into a Data Governance Control Board (Figure 6) with each working group responsible for a particular data domain / data entity (e.g. Customer) or a data subject area that consists of multiple data entities.

Data governance organization structure

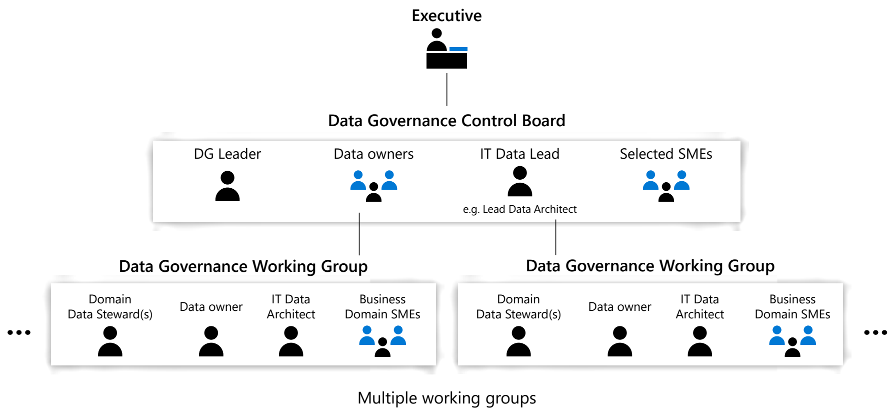
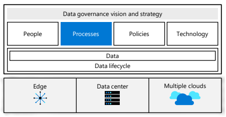


Figure 6

Data Governance Processes



There are four categories of data governance processes shown below:

Process Category	Processes
Data discovery processes (to understand the data landscape)	• A data and data entity discovery, mapping and cataloguing process • A data profiling discovery process to determine the quality of data • A sensitive data discovery and governance classification process • A data maintenance discovery process for CRUD[3] analysis (e.g. from log files) to understand usage and maintenance of data (e.g. master data) across the enterprise
Data governance definition processes	• Create and maintain a common business vocabulary in a business glossary. This involves defining data entities (including master data), data attributes names, data integrity rules and valid formats • Define reference data to standardise code sets across the enterprise • Define data governance classifications schemes to label data to determine how to govern it • Define data governance policies and rules to govern data entity and document lifecycles • Define success metrics and threshold
Data governance policy and rule enforcement processes	• A process to automate application / enforcement of data governance policies and rules • A process to manually apply and enforce policies and rules • Event-driven, on-demand and timer-driven (batch) data governance processes published as services that can be invoked to govern: • Data ingestion - cataloguing, classification, owner assignment, and storing • Data quality • Data access security • Data privacy • Data usage e.g. including sharing and to ensure licensed data is only used for approved purposes • Data maintenance e.g. of master data • Data retention • Master data and reference data synchronisation
Monitoring processes	• Monitor and audit data usage activity, data quality, data access security, data privacy, data maintenance and data retention • Monitor policy rule violation detection and resolution

[3] CRUD = Create, Read, Update, Delete

Governance processes set out how to discover, define, enforce and monitor

The common business vocabulary should be defined in a business glossary within a data catalog. Following on from the discussion on data governance working groups, each working group should take responsibility for defining a specific data entity or data subject area (multiple related entities). Therefore, multiple data entities in the vocabulary, along with the policies and rules, can be worked on in parallel (see Figure 7).

The common business vocabulary should be defined in the business glossary within the data catalog

Policies and rules should also be defined in the catalog to govern data quality, data privacy, data access security and retention across the distributed data landscape

Defining data entities in a common business vocabulary

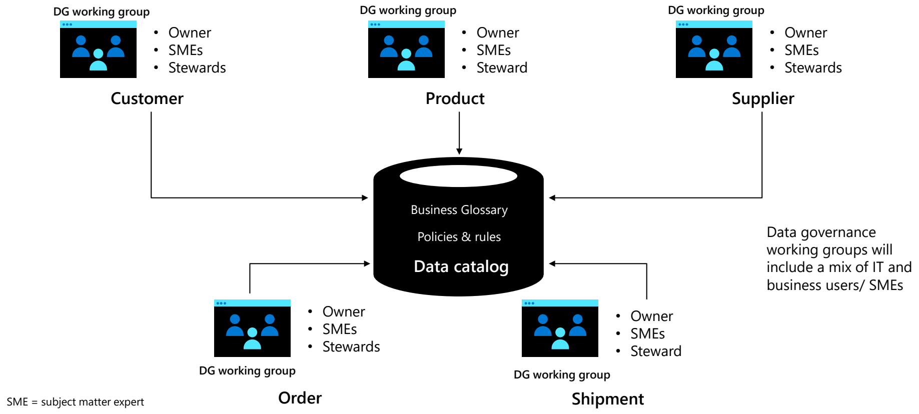


Figure 7

Integration of the catalog business glossary with other technologies such as ETL tools, data modelling tools, BI tools, DBMSs, master data management, data virtualisation tools, software development tools etc.., is then needed to get consistent common data names into all technologies.

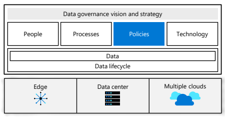
A data concept model is a good top down approach to identify data entities to get your common business vocabulary started

Different data governance working groups can work on different data entities using automatic data discovery in the catalog to identify attributes to add to the common business vocabulary to describe those data entities

A good practice to quick start the creation of a common business vocabulary is to create a data concept model. This top down approach gets you started because it identifies data concepts that can be used as data entities in a common business vocabulary. It is then possible to assign a different data governance working group to each data concept (entity) or group of related data concepts (subject area). In this way different working groups are assigned to govern different data entities across the landscape. During the build of a common business vocabulary, it should be possible to use data catalog software to automatically discover what data is out there across multiple data stores to help identify all the attributes associated with specific data entities. This is a bottom-up approach. By using a top down approach of a data concept model to get you started and a bottom up automated data discovery approach to identify the attributes of a data entity, it should be possible for multiple working groups to incrementally build up a common business vocabulary reasonably quickly.

Using a data catalog for automated data discovery enables the mapping of disparate data to a common vocabulary to understand where the data for each particular data entity in the business glossary is actually located across the enterprise.

Policies and Rules to Govern Data at Different Points in The Lifecycle



Policies and rules need to be defined to improve data quality, protect sensitive data and govern retention

Data governance policies describe a set of rules to control the integrity, quality, access security, privacy and retention of data. There are different types of policy including:

- Data integrity policies e.g. valid values, referential integrity
- Data quality policies with data standardisation, cleansing and matching rules
- Data protection policies with access security and data privacy rules
- Data retention policies to manage the lifecycle with retention, archive and backup rules

Note that multiple versions of a policy may be needed to govern the same data across different legal jurisdictions.

Looking back to the data governance classification schemes in Figure 2, the data confidentiality classification scheme has five classification levels. These are Public, Internal Use Only, Confidential, Sensitive PII and Restricted.

Policies and rules should be created for each class in a classification scheme

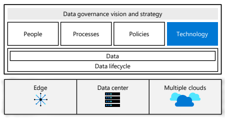
Each class in a data governance classification scheme should be used as a tag to label data to say how it should be governed

The way to govern data is to combine this data governance classification scheme with policies and rules. So, for example, consider each of the five levels as a label that can be used to label data. Take for example 'Sensitive PII'. By creating rules for Sensitive PII data and attach these rules to a policy you create a policy for Sensitive PII data. You can then attach the policy to the Sensitive PII label and then attach the Sensitive PII label to the data. In this way all data labelled as Sensitive PII is subject to the same policies and rules. This is known as tag-based policy management. It is flexible because an individual rule or a policy can be independently changed, and all data labelled Sensitive PII would then be governed by the new rules. Equally, a Sensitive PII label can be detached from data and a Confidential label used instead. In this case the data instantly becomes governed by a new set of policies and rules associated with the Confidential label.

Once policies and rules are defined in a data catalog for each class in a data governance classification scheme, they can be passed to other technologies from a data catalog (via APIs) for them to enforce. Alternatively, a common data management platform (data fabric) that can connect to multiple data stores could potentially enforce them.

It should then be possible to monitor data quality, privacy, access security, usage, maintenance and retention of specific data entities through their lifecycle.

Data Governance Technology



The technologies needed for data governance are:

- A data catalog that includes:
  - A business glossary
  - Automated data discovery, profiling, tagging, cataloguing and mapping to a glossary
  - Automated sensitive data detection and governance classification
  - Interoperability with other catalogs, tools and applications to share metadata via APIs and open standards

A data catalog, data fabric software, a data lake and master data management are all key technologies needed to help govern data and create trusted data assets

- A data lake to ingest and process data
- Enterprise data fabric software with built-in support for:
  - Data centre, multi-cloud and edge data connectivity
  - Data stewardship tooling
  - Data cleansing and integration
  - Metadata lineage
  - Data privacy masking
  - Universal data access security across multiple data stores in a distributed data landscape
- Data stores that support data encryption, dynamic data masking and integration with the data catalog
- AI assisted data governance
- Master and reference data management

Technology Needed For End-To-End Data Governance

Enterprise data catalog, and Azure Data Factory are key technologies to help you govern data

In the context of technology needed for end-to-end data governance, Microsoft provides its own technologies and also partner technologies on Azure.

Microsoft provides the following technology components to assist you in governing data:

- Microsoft Common Data Model
- Azure Data Lake Storage
- Azure Data Factory

Microsoft Common Data Model

Microsoft has created an open common data model to describe core data entities that need to be shared across the enterprise

The first step in data governance is to create a common business vocabulary of common data names and definitions describing logical data entities that can be shared across the enterprise. For example, customer, account, product, supplier, orders, payments, returns etc. Once this has been done, it then becomes possible to create these common data assets and store them where their reuse can be maximised to drive consistency everywhere.

The Microsoft Common Data Model (CDM) is an open, pre-built set of common business entities and activities used across a business that can be used to shortcut the creation of your common business vocabulary.

Microsoft CDM can be used to 'quick start' your common business vocabulary.

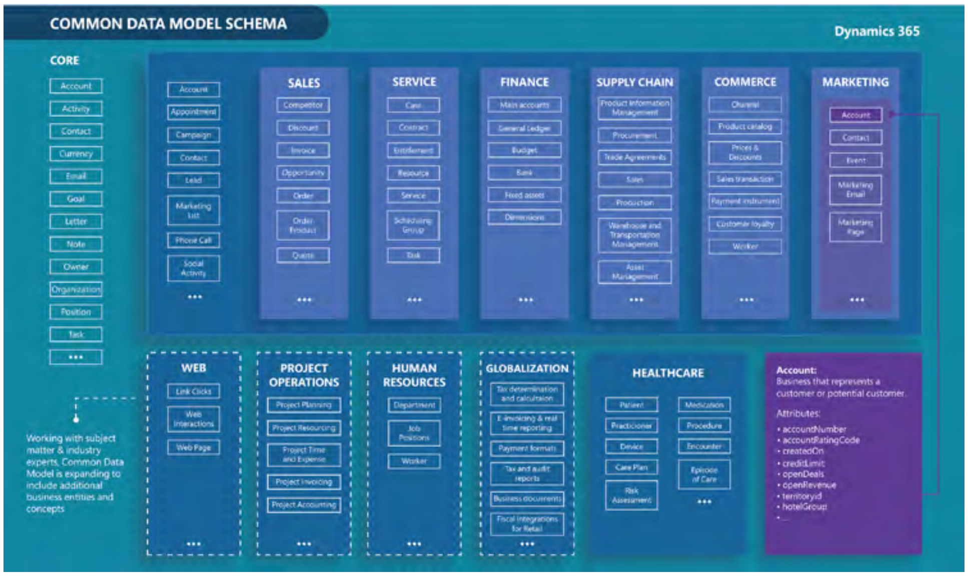


Figure 8

Azure Data Lake Storage

Azure Data Lake Storage is shared storage that underpins Microsoft Azure Synapse Analytics, Azure ML, Azure Databricks and Azure HD Insight

ADLS is also accessible by Power BI

Azure Data Lake Storage (ADLS) provides a common place to capture / ingest and integrate data to produce trusted data assets. CDM entities can be created in Azure Data Lake storage that is accessible to Power BI, Azure Data Factory, Azure Databricks, Azure Synapse Analytics and Azure ML. See Figure 10 below.

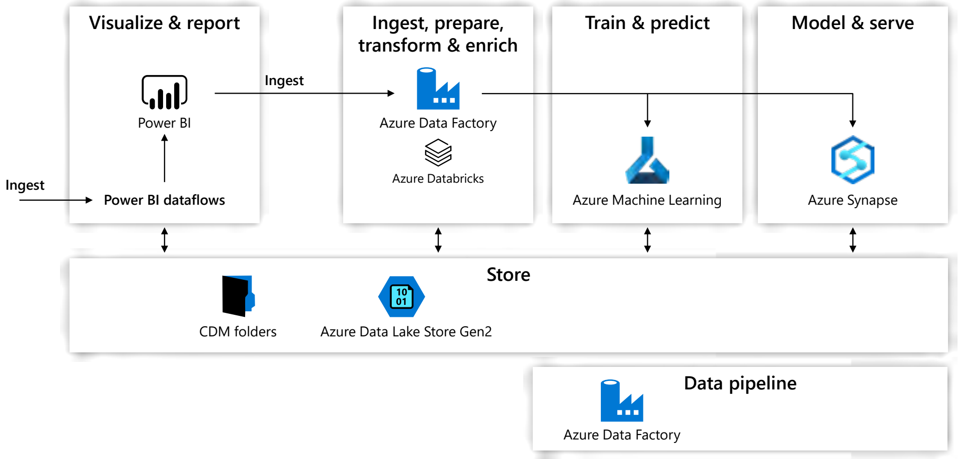


Figure 10

Microsoft Azure Data Factory (ADF)

Microsoft's strategic pay-as-you-use data management platform (data fabric) for cleaning and integrating data is Azure Data Factory (ADF)

ADF allows you to build scalable data integration pipelines code free

Microsoft Azure Data Factory is a fully managed, pay-as-you-use, hybrid data integration service for highly scalable ETL and ELT processing. It uses Spark to process and analyse data in parallel and in memory to maximise throughput.

It supports over 80 connectors to external data sources and databases and has templates for common data integration tasks. A visual front-end browser-based GUI enables non-programmers to create and run process pipelines to ingest, transform and load data, while more experienced programmers have the option to incorporate custom code if required (e.g. Python programs).

ADF enables collaborative development between business and IT professionals in the creation of reusable trusted data assets

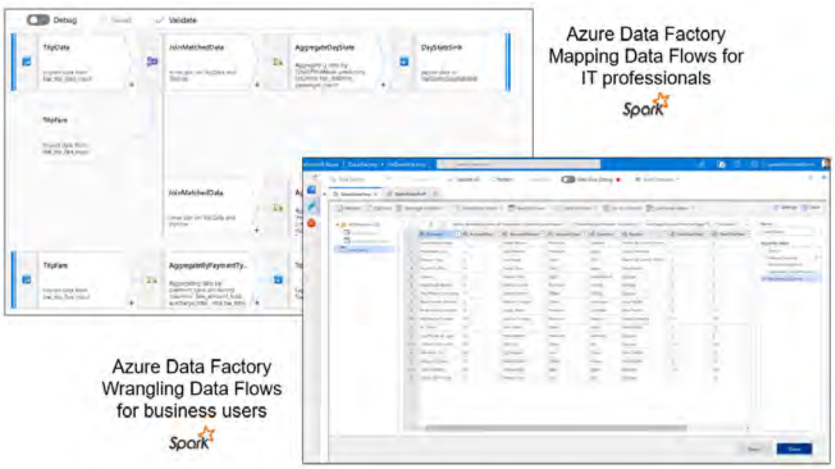

Debug
Validate
TripData
JoinMatchedData
AggregateDayStats
DayStatsSink
TripFare
Azure Data Factory Mapping Data Flows for IT professionals
Spark
Azure Data Factory Wrangling Data Flows for business users
Spark

Figure 11

Development of simple or comprehensive ETL and ELT processes without coding or maintenance, including ingest, move, prepare, transform and process your data can be achieved with a few clicks. Scheduling and triggers can also be designed and managed within Azure Data Factory to build an automated data integration and loading environment for producing trusted data assets that are described in the Azure Data Catalog business glossary.

ADF can be used to implement and manage a hybrid environment, which includes connectivity to on-premise, cloud, edge streaming and SaaS data (e.g. from applications such as Salesforce), in a secure and consistent way.

ADF wrangling data flows enables business users to make use of the platform to visually discover, explore and prepare data at scale without writing code. This easy to use ADF capability is similar to Microsoft Excel Power Query or Microsoft Power BI Dataflows where business users use a spreadsheet style user interface with drop-down transforms to prepare and integrate data.

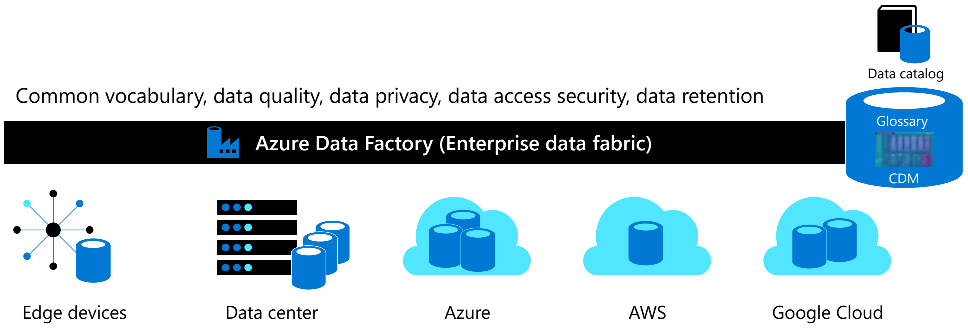
Combining Microsoft Technologies to Help Govern Data

In the context of data governance, these technologies can be combined to produce trusted reusable data assets. This is shown in Figure 12 and 13.

Data in disparate registered data sources across the data landscape can be ingested into Azure Data Lake Storage and integrated using Azure Data Factory to create trusted, commonly understood, reusable CDM data assets that can be persisted back in the data lake published in Azure Data Catalog

Everything that is underpinned by ADLS in Figure 10 can then make use of trusted, commonly understood CDM described data assets

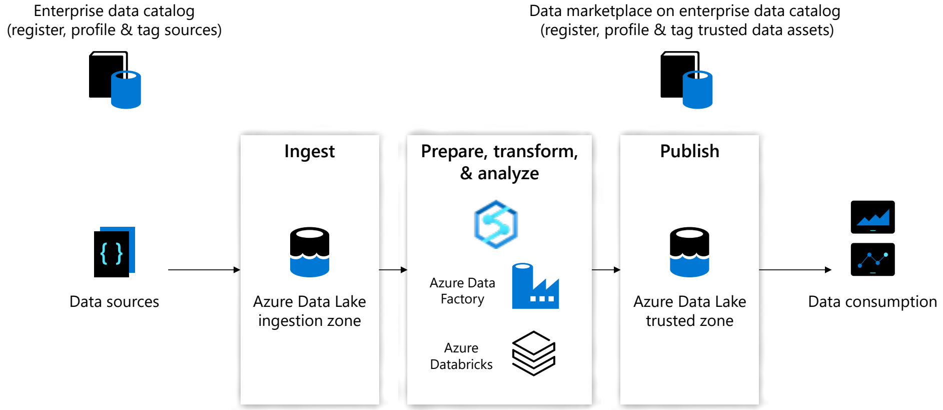
The objective is build once, publish in a data marketplace (Azure Data Catalog) and reuse everywhere



A common data fabric

Data catalog
Common vocabulary, data quality, data privacy, data access security, data retention
Azure Data Factory (Enterprise data fabric)
Glossary
CDM
Edge devices
Data center
Azure
AWS
Google Cloud

Figure 12



Data cataloging

Enterprise data catalog
(register, profile & tag sources)
Data marketplace on enterprise data catalog
(register, profile & tag trusted data assets)
Ingest
Prepare, transform, & analyze
Publish
Data sources
Azure Data Lake ingestion zone
Azure Data Factory
Azure Databricks
Azure Data Lake trusted zone
Data consumption

Figure 13

Microsoft Partner Technologies for Data Governance

Microsoft partners also offer technology on Azure to help with data governance

Azure Marketplace Partners for data governance include:

- Tamr (for ETL processing)
- Talend (for ETL processing and Data Cataloguing)
- Informatica (for Enterprise Data Cataloguing)
- Qlik Data Integration
- Semarchy (master data management)
- Profisee (master data management)

Managing Master Data

A master data management (MDM) system is a core component needed in data governance

Master data entities can be defined in the Azure Data Catalog business glossary as part of a common business vocabulary

Disparate master data can then be ingested into ADLS from where it can be cleaned, matched and integrated using ADF to populate an MDM system

Master data maintenance also needs to be governed

Central to any data governance program is master data management. Creating trusted master data is therefore critical. This can be done by defining master data entities in the business glossary within Azure Data Catalog and then using the data catalog to register data sources and discover where disparate master data is located across multiple data stores in the distributed data landscape.

By mapping the physical data names of discovered disparate master data to the common business vocabulary in Azure Data Catalog, it then becomes possible to know how to clean, match and integrate the data discovered to create golden master data records stored in a central MDM system. This can be done using Azure Data Lake Storage and Azure Data Factory as shown in Figure 13. Once created and stored centrally, master data can then be synchronised with all other systems that need it to make sure they are consistent.

In addition, master data maintenance needs to be governed. The challenge is therefore to identify in which tasks of which business processes that maintenance occurs. This can be done using business process identification and CRUD analysis. However, it is often a manual task to work this out but is now helped by the emergence of process mining and analysing database log files. Once the tasks within a process that maintain master data have been identified, it can be governed.

Governing GDPR Consent Management Using Master Data

GDPR consents can be stored along side customer master data to govern customer data usage

Finally, master data management (MDM) provides the ideal place for GDPR customer consent management. This can be done by collecting consents from all applications that request it, matching these with customer master data and storing all consents in additional tables along with the master data record in the MDM system.

Data Governance Maturity Model

Looking at the data governance challenge, you may be wondering how mature you are in terms of covering all aspects of this across your data landscape. In order to assess that, the following data governance maturity model is provided.

	Ungoverned	Stage 1	Stage 2	Fully governed
People	No stakeholder executive sponsor	Stakeholder sponsor in place	Stakeholder sponsor in place	Stakeholder sponsor in place
	No roles and responsibilities defined	Roles and responsibilities defined	Roles and responsibilities defined	Roles and responsibilities defined
	No DG control board	DG control board in place but no ability	DG control board in place with data	DG control board in place with data
	No DG working groups	No DG working groups	Some DG working groups in place	All DG working groups in place
	No data owners accountable for data	No data owners accountable for data	Some data owners in place	All data owners in place
	No data stewards appointed with responsibility for data quality	Some data stewards in place for DQ but scope too broad e.g. whole dept	Data stewards in place and assigned to DG working groups for specific data	Data stewards in place assigned to DG working groups for specific data
	No one accountable for data privacy	No one accountable for data privacy	CPO accountable for privacy (no tools)	CPO accountable for privacy with tools
	No one accountable for access security	IT accountable for access security	IT Sec accountable for access security	IT Sec accountable for access security & responsible for enforcing privacy
	No one to produce trusted data assets	Data publisher identified and accountable for producing trusted data	Data publisher identified and accountable for producing trusted data	Data publisher identified and accountable for producing trusted data
	No SMEs identified for data entities	Some SMEs identified but not engaged	SMEs identified & in DG working groups	SMEs identified & in DG working groups
Process	No common business vocabulary	Common biz vocabulary started in a glossary	Common business vocabulary established	Common business vocabulary complete
	No way to know where data is located, its data quality or if it is sensitive data	Data catalog auto data discovery, profiling & sensitive data detection on some systems	Data catalog auto data discovery, profiling & sensitive data detection on all structured data	Data catalog auto data discovery, profiling & sensitive data detection on structured & unstructured in all systems w/ full auto tagging
	No process to govern authoring or maintenance of policies and rules	Governance of data access security policy authoring & maintenance on some systems	Governance of data access security, privacy & retention policy authoring & maintenance	Governance of data access security, privacy & retention policy authoring & maintenance
	No way to enforce policies & rules	Piecemeal enforcement of data access security policies & rules across systems with no catalog integration	Enforcement of data access security and privacy policies and rules across systems with catalog integration	Enforcement of data access security, privacy & retention policies and rules across all systems
	No processes to monitor data quality, data privacy or data access security	Some ability to monitor data quality Some ability to monitor privacy (e.g. queries)	Monitoring and stewardship of DQ & data privacy on core systems with DBMS masking	Monitoring and stewardship of DQ & data privacy on all systems with dynamic masking
	No availability of fully trusted data assets	Dev started on a small set of trusted data assets using data fabric software	Several core trusted data assets created using data fabric	Continuous delivery of trusted data assets with enterprise data marketplace
	No way to know if a policy violation occurred or process to act if it did	Data access security violation detection in some systems	Data access security violation detection in all systems	Data access security violation detection in all systems
	No vulnerability testing process	Limited vulnerability testing process	Vulnerability testing process on all systems	Vulnerability testing process on all systems
	No common process for master data creation, maintenance & sync	MDM with common master data CRUD & sync processes for single entity	MDM with common master data CRUD & sync processes for some data entities	MDM with common master data CRUD & sync processes for all master data entities complete

Benchmark your company on this data governance maturity model to gauge your progress

Policies	No data governance classification schemes on confidentiality & retention	Data governance classification scheme for confidentiality	Data governance classification scheme for both confidentiality and retention	Data governance classification scheme for both confidentiality and retention
	No policies & rules to govern data quality	Policies & rules to govern data quality started in common vocabulary in business glossary	Policies & rules to govern data quality defined in common vocabulary in catalog biz glossary	Policies & rules to govern data quality defined in common vocabulary in catalog biz glossary
	No policies & rules to govern data access security	Some policies & rules to govern data access security created in different technologies	Policies & rules to govern data access security & data privacy consolidated in the data catalog using classification scheme	Policies & rules to govern data access security, data privacy and retention consolidated in the data catalog using classification schemes and enforced everywhere
	No policies & rules to govern data privacy	Some policies & rules to govern data privacy	Policies & rules to govern data access security & data privacy consolidated in the data catalog using classification scheme	Policies & rules to govern data access security, data privacy and retention consolidated in the data catalog using classification schemes and enforced everywhere
	No policies & rules to govern data retention	No policies & rules to govern data retention	Some policies & rules to govern data retention	Policies & rules to govern data access security, data privacy and retention consolidated in the data catalog using classification schemes and enforced everywhere
	No policies & rules to govern master data maintenance	Policies & rules to govern master data maintenance for a single master data entity	Policies & rules to govern master data maintenance for some master data entities	Policies & rules to govern master data maintenance for all master data entities
Technology	No data catalog with auto data discovery, profiling & sensitive data detection	Data catalog with auto data discovery, profiling & sensitive data detection purchased	Data catalog with auto data discovery, profiling & sensitive data detection purchased	Data catalog with auto data discovery, profiling & sensitive data detection purchased
	No data fabric software with multi-cloud edge and data centre connectivity	Data fabric software with multi-cloud edge and data centre connectivity & catalog integration purchased	Data fabric software with multi-cloud edge and data centre connectivity & catalog integration purchased	Data fabric software with multi-cloud edge and data centre connectivity & catalog integration purchased
	No metadata lineage	Metadata lineage available in data catalog on trusted assets being developed using fabric	Metadata lineage available in data catalog on trusted assets being developed using fabric	Metadata lineage available in data catalog on trusted assets being developed using fabric
	No data stewardship tools	Data stewardship tools available as part of the data fabric software	Data stewardship tools available as part of the data fabric software	Data stewardship tools available as part of the data fabric software
	No data access security tool	Data access security in multiple technologies	Data access security in multiple technologies	Data access security enforced in all systems
	No data privacy enforcement software	No data privacy enforcement software	Data privacy enforcement in some DBMSs	Data privacy enforcement in all data stores
	No master data management system	Single entity master data management system	Multi-entity master data management system	Multi-entity master data management system

Conclusions

Benchmark your company on this data governance maturity model to gauge your progress

The key to successful data governance is to break structured data down into data entities and data subject areas and then make use of a data governance solution to surround specific data entities and data subject areas with people, processes, policies and technology to govern the lifecycle of each of those data entities. This can be done by establishing a common business vocabulary in a business glossary within a data catalog.

The data catalog is critical to success

The data catalog is critical technology because you cannot govern data if you don’t know where it is or what it means. Data catalog software provides automatic data discovery, automatic profiling to determine its quality and automatic sensitive data detection. In addition, it helps map disparate data to your common vocabulary data names and definitions in the catalog business glossary to understand what data means.

Confidentiality and retention data governance classification schemes guide the creation of policies and rules to govern data

Creating data governance classification schemes such as the examples shown in Figure 2 provide different levels of governance classification. These need to be defined in the data catalog. At this point, policies and rules can then be created in the data catalog and associated with different levels of governance classification.

It should then be possible to label (or tag) data attributes in the business glossary with confidentiality and retention classes to specify how to govern it. And because the data catalog *already knows* the mappings of physical data attributes in different data stores to attributes in business glossary, then labelling an attribute in the glossary automatically determines how to govern data mapped to it in underlying data stores. Multiple technologies that integrate with the data catalog can then access this metadata to consistently enforce these policies and rules across all data stores in a distributed data landscape. The exact same governance classification labels can also be applied to unstructured data.

MDM is also important because master data is so widely shared across both operational transaction processing systems and analytical systems

Master data entities are critical because this data is so widely shared. It is also frequently associated with documents. For example, a customer and an invoice, a supplier and a contract, an asset and an operating manual. Therefore, master data values (e.g. supplier name) can be used to tag related documents to ensure that relationships between structured and unstructured data are preserved.

Creating trusted, reusable data assets descibed using a common business vocabulary and published in a data marketplace enables trusted data to be widely shared

Data governance helps to systematically create trusted, and protected data

Using the common vocabulary data entities defined in the data catalog, and the mappings discovered, it should then be possible to create pipelines using data fabric to create trusted data assets that can be published in a data marketplace for all to share. The key point about data governance is that there are methods here to get your data under control and once trusted, to then use it to drive value. Success will be determined by how well you organise and collaborate to do it. This Microsoft Data Governance Guide is provided to assist with that so that you can systematically make use of people, processes, policies and technology to get your data into a trusted well governed state to eradicate data quality problems and the impact they have, uphold privacy, secure access and drive business value.