WHITE PAPER

INTELLIGENT
BUSINESS
STRATEGIES

The Value of Data Virtualisation in a Data Mesh

By Mike Ferguson
Intelligent Business Strategies
January 2022

Research sponsored by:
denodo

Table of Contents

WHAT IS DATA MESH?

Data Mesh is a decentralised approach to producing trusted, reusable datasets know as data products

Data Mesh is a data architecture concept first defined in an article[1] published two and a half years ago. Since then, interest in Data Mesh has grown rapidly. based on a decentralised business domain-oriented approach to producing trusted, reusable datasets known as “data products” to share and consume across the enterprise. The intention is to use these data products primarily in analytical environments.

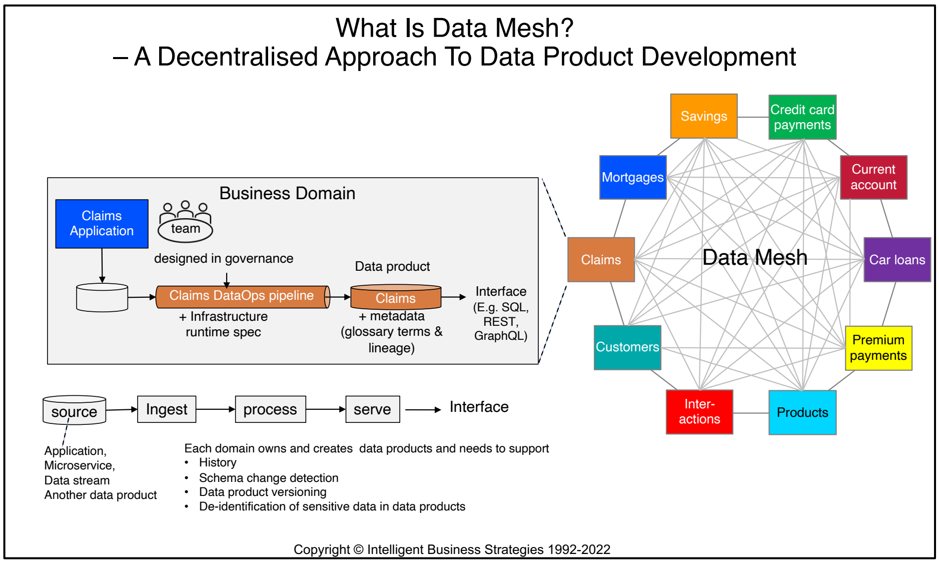
The Data Mesh concept is based on four major principles:

- Domain oriented decentralised data ownership and architecture
- Data as a product
- Self-serve data infrastructure as a platform
- Federated computational data governance

Business domain subject matter experts use self-service tools to create pipelines that produce data products

The idea behind Data Mesh is that people in business domains who work with specific data every day utilise self-service infrastructure software to create pipelines that take data from application data sources used in that busines domain and produce data products that are available in a Data Mesh.

Each domain owns and creates data products that can be consumed elsewhere in the enterprise


What Is Data Mesh?
– A Decentralised Approach To Data Product Development

Each domain owns and creates data products and needs to support

- History
- Schema change detection
- Data product versioning
- De-identification of sensitive data in data products

Copyright © Intelligent Business Strategies 1992-2022

Figure 1

A key objective in this approach is to speed up the creation and availability of trusted, high-quality, compliant data to share across the enterprise by training business professionals to produce this data rather than always relying on centralised IT teams who may not be able to keep pace with demand. IT professionals can be embedded in business domains to help achieve this.

WHAT PROBLEM IS DATA MESH TRYING TO ADDRESS?

Many new data sources have emerged that businesses want to analyse

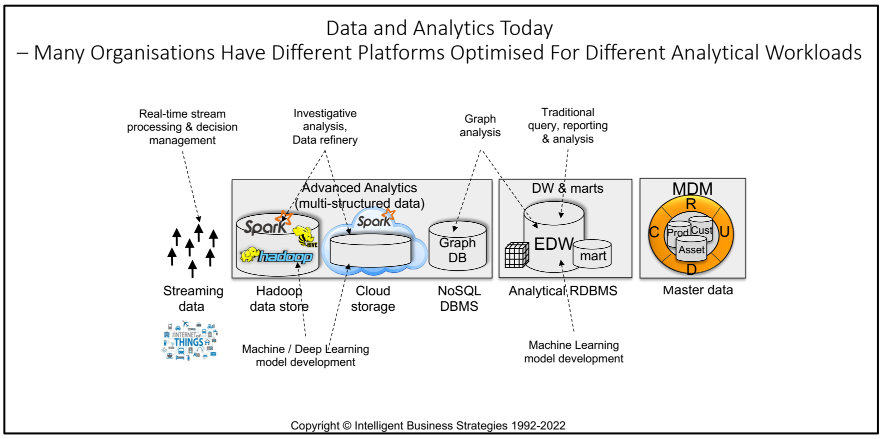
One question worth asking is what problem is Data Mesh trying to address? To explain this, we need to look at what has and is happening with data.

[1] How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh, Zhamak Dehghani, May 2019

Increasingly, many new data sources have emerged with data that companies now want to analyse. This includes traditional structured data in transaction databases, machine generated data such as clickstream data, infrastructure log data and IoT data as well as human generated data such as inbound emails, web chat, voice and social network data.

Many companies now have different analytical workloads analysing overlapping subsets of data running on different centralised analytical systems

All of this data has and is being created and ingested into multiple different applications and data stores that are both on-premises and in multiple clouds. Data is also streaming in off IoT devices at the edge. The result is that many companies now have different analytical workloads running on different analytical systems (both on premises and in the cloud) analysing overlapping subsets of this data (see Figure 2).


Data and Analytics Today
– Many Organisations Have Different Platforms Optimised For Different Analytical Workloads
Real-time stream processing & decision management
Investigative analysis, Data refinery
Graph analysis
Traditional query, reporting & analysis
Advanced Analytics (multi-structured data)
Spark
hadoop
Spark
Graph DB
DW & marts
EDW
mart
MDM
R
C
U
D
Prod
Cust
Asset
Streaming data
Hadoop data store
Cloud storage
NoSQL DBMS
Analytical RDBMS
Master data
Machine / Deep Learning model development
Machine Learning model development

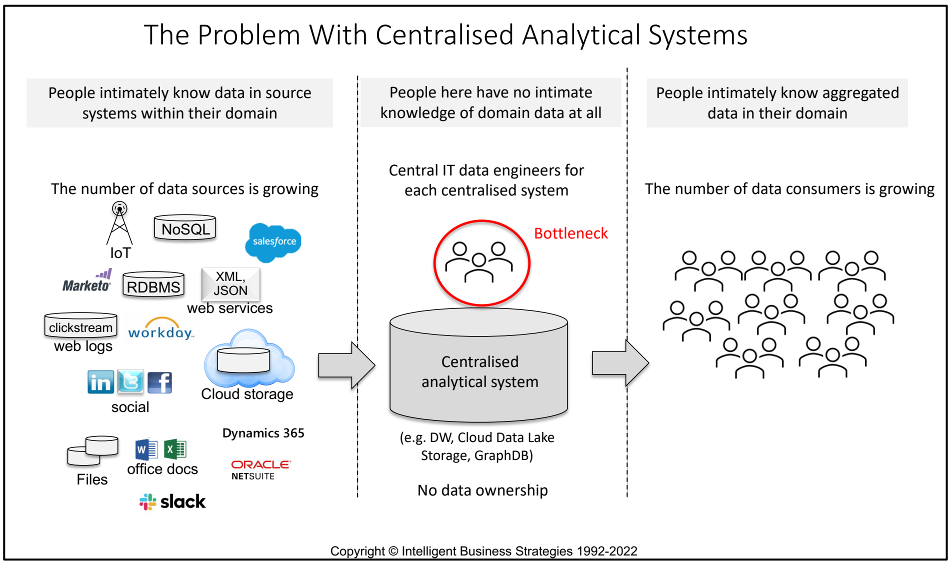
Copyright © Intelligent Business Strategies 1992-2022

Figure 2

As the number of data sources grows, all data flowing to centralised analytical systems is not guaranteed

Also centralised IT-based data engineers can't keep pace with demand and so become a bottleneck

These analytical systems are centralised which means that they require data to be brought to a central system before the data is cleaned, transformed, integrated, and analysed. However, as the number of new data sources grows, the chances of all data flowing to a central analytical system is not guaranteed. In addition, centralised data engineers, whether they be IT professionals or data scientists are unlikely to be able to keep pace with business demands and are already being viewed as a bottleneck in many cases. In addition, centralised data engineers often have little in the way of intimate knowledge of domain data sources. This is shown in Figure 3.


The Problem With Centralised Analytical Systems
People intimately know data in source systems within their domain
People here have no intimate knowledge of domain data at all
People intimately know aggregated data in their domain
The number of data sources is growing
Central IT data engineers for each centralised system
The number of data consumers is growing
IoT
NoSQL
salesforce
Marketo
RDBMS
XML, JSON
web services
clickstream
web logs
workday.
social
Cloud storage
Dynamics 365
Files
office docs
ORACLE NETSUITE
slack
Bottleneck
Centralised analytical system
(e.g. DW, Cloud Data Lake Storage, GraphDB)
No data ownership

Copyright © Intelligent Business Strategies 1992-2022

Figure 3

Each different analytical system is operating as an independent silo

In addition, the centralised analytical systems in Figure 1 are siloed and typically operate independently with the same data often repeatedly extracted, cleaned, transformed, and integrated for each different analytical workload in each silo (Figure 4). A good example might be customer data, product data or orders data.

The same data is often repeatedly extracted transformed and integrated for different analytical workloads in different systems

Siloed analytical systems increase data integration costs and cause reinvention which can lead to inconsistent data

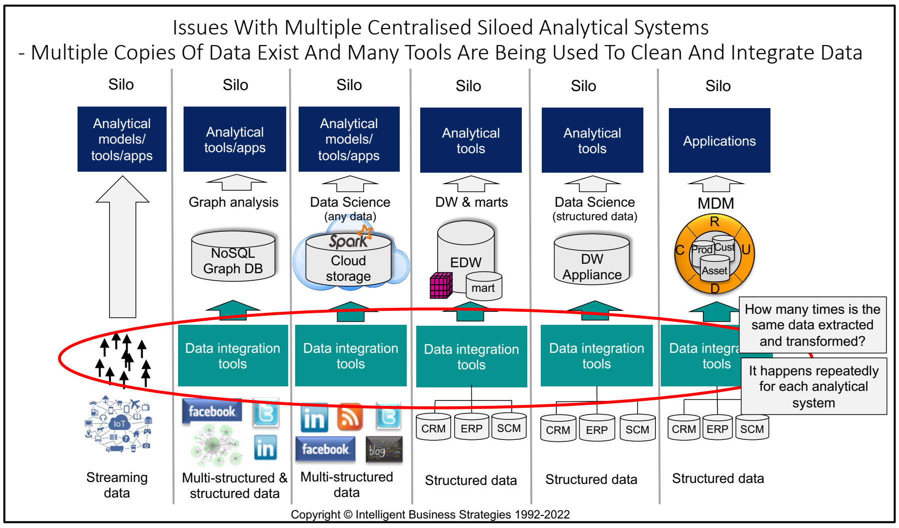


Figure 4

It would be quicker and less costly to do it once and create reusable 'data products' for reuse in different analytical systems

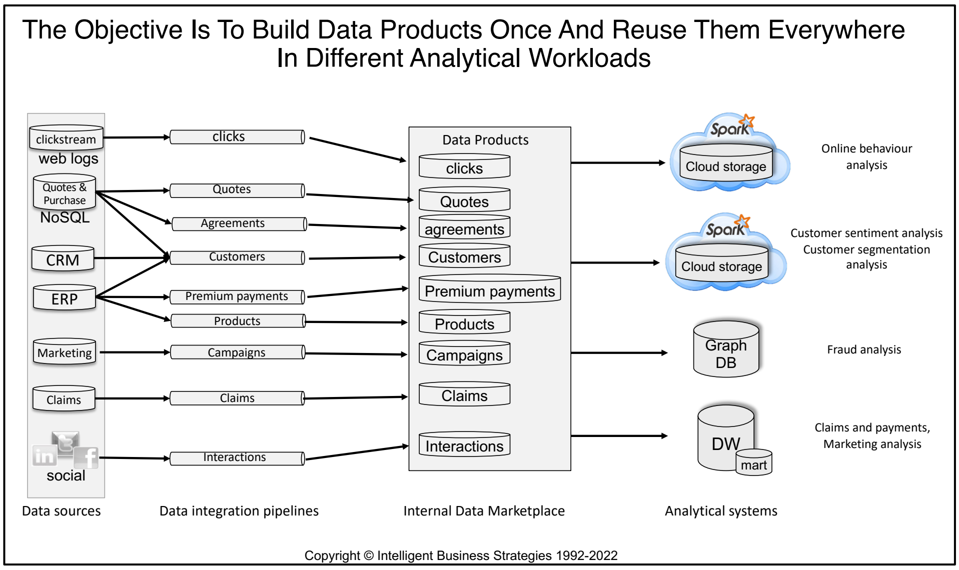
Ideally it would be better to build trusted, compliant "data products" once and reuse them in different analytical workloads rather than repeatedly re-inventing data integration pipelines to create that same data for each different analytical system. Data Mesh is intended to solve this problem while also speeding up the rate at which data products can be created by enabling different teams in different business domains to produce historical, compliant data for consumption in different analytical systems and workloads being implemented in other parts of the enterprise.

What Are Data Products?

Data products are reusable data sets that can be consumed by different analytical systems in support of different analytical workloads

At the heart of a Data Mesh are data products. An Insurance example of this is shown in Figure 5 where data integration pipelines are created to produce data products. These data products are based on business data concepts like Customers, Products, Agreements (policies), Premium Payments, Claims etc. Data products produced are then published in a data marketplace (a catalog) to make them available for consumption by different analytical systems that support different types of analytical workloads. This includes model development in data science, data warehouses providing business intelligence and graph databases supporting graph analysis e.g., for Fraud. The idea is to 'build once and reuse everywhere'.

Build once, reuse everywhere


The Objective Is To Build Data Products Once And Reuse Them Everywhere In Different Analytical Workloads

Copyright © Intelligent Business Strategies 1992-2022

Figure 5

Data products should be high quality, compliant, and secure and be published in a data marketplace with metadata and APIs to make them easy find, consume and use

Data products are defined[2] as having the following characteristics:

- Discoverable
- Addressable
- Trustworthy
- Self-describing
- Interoperable
- Secure

In addition, the data product includes the data itself, the pipeline (rules to clean and integrate data), the runtime specification to execute the pipeline and APIs.

Types of Data Product in a Data Mesh

Different types of data products can exist in a Data Mesh. This includes:

- Physical data products
- Virtual data products

[2] How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh, Zhamak Dehghani, May 2019

- Stored queries

Physical data products

Physical data products are persisted datasets that have been produced, stored, and published to make them available for consumption.

Data products can be persisted, virtual or stored queries

Virtual data products

Virtual data products can be defined using data virtualisation software. They are virtual views that integrate data on-demand from one or more underlying data sources. These virtual views can be queried and can be published in a data catalog for people to discover and use. In addition, a virtual data product can be created to provision all or some of the data from one or more underlying physical data products and other virtual data products.

Stored queries as a service

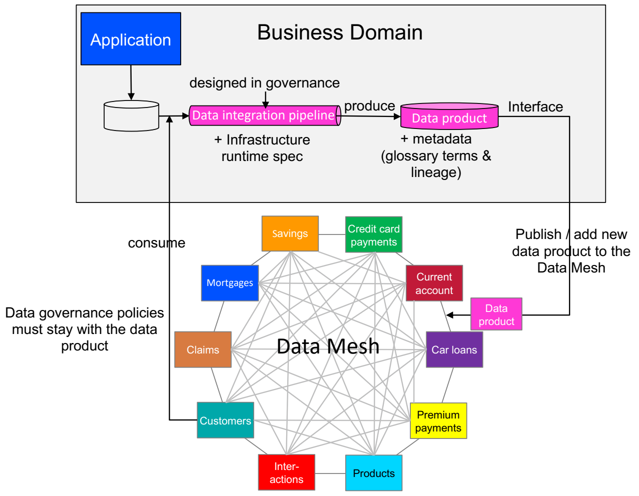
Stored queries are typically SQL queries or SQL scripts that can be published as services and invoked on demand e.g., via a REST API or GraphQL. When executed, the stored query will then produce a data product and serve it. Stored queries can potentially integrate data from one or more data sources including other data products. They can query virtual data products for example.

CONSUMING DATA PRODUCTS AND PRODUCING NEW ONES IN A DATA MESH

It is also possible to consume data products as part of a pipeline that produces a new data product. This can be done by consuming any type of data product whether it be physical, virtual, or stored query. An example of this is shown in Figure 6.

Data products can be consumed and used in the creation of new data products that can themselves be added to a Data Mesh

Domains Can Consume Data Products Produced By Other Domains And Then Create New Data Products To Add To A Data Mesh



Copyright © Intelligent Business Strategies 1992-2022

Figure 6

Build on what has already been created rather than reinvent it

This approach encourages reuse and allows organisations to see and build on what has already been created as opposed to re-inventing it. The benefits are that it reduces cost of data integration across organisations and shortens the time to value through reuse of data products. It also improves data governance by ensuring reuse of high quality data and allows data access security policies to be defined on data product once and have them apply wherever that data is accessed, consumed and used.

Critical Capabilities for Data Mesh to Succeed

Having outlined the problem that Data Mesh is trying address, there are a number of capabilities that need to be available to enable this decentralised approach to development of data products to succeed. These are as follows:

A Business Glossary

A business glossary provides a business lens on data

A business glossary is typically held within a data catalog and provides a business lens on data. The business glossary can be used to explain:

- The meaning of raw data in data sources to data producers
- The meaning of data in data products to consumers
- Who the owners of data products are

A business glossary enables consumers to clearly understand the meaning of data in each data product

The whole purpose of Data Mesh is to enable data products to be created that can be easily shared across the enterprise in support of different analytical initiatives. People must be able to understand the business meaning of data available to make this possible. Therefore, a common business vocabulary of business data names and descriptions needs to be created in a business glossary to describe the meaning of data attributes within each data product. Data consumers can then refer to the glossary if they are unsure of data meaning which should give them the confidence to use it.

Data Catalog

A data catalog helps producers understand what data is available and what is sensitive. It also helps consumers understand how data products have been produced

In addition to a business glossary, a data catalog provides automatic data discovery and profiling of data in data sources across a distributed data landscape. It also provides automatic detection and classification of all sensitive data types. This is needed to help domain-based producers of data products to understand the quality and sensitivity of data so that they can produce high-quality compliant data products. Data catalogs also provide data lineage to provide clear understanding of how each data product has been produced. This includes the data sources of a data product, visibility of transformation and integration steps and any dependencies e.g., if another data product is used.

Common Data Infrastructure Software

Using common software to create data products enables metadata to be easily shared across business domains

A key reason common data infrastructure software is needed, is to standardise the software used by domain-based teams to produce data products. It also enables metadata to be easily shared across all business domains who are involved in producing data products. This includes metadata on data products (glossary and lineage), sensitive data, and data privacy and access security policies to govern data and allow it to remain protected when shared.

Enterprise Data Marketplace

An enterprise data marketplace allows consumers to easily see what data products exist

It also manages the provisioning and auditing of data products for consumption and use

The purpose of a data marketplace is to enable consumers to easily find out what business ready data products exist, see who the data product owners are, see common definitions for data products in a business glossary and lineage on how each product was created. A data marketplace also governs data product access security and ensures compliant data provisioning for those who want to consume it. Business consumer ratings, auditing of data product consumption and traceability of data product usage is also needed. A data catalog of business ready data products can fulfil the role of a data marketplace.

THE ROLE OF DATA VIRTUALISATION IN A DATA MESH

Having looked at what Data Mesh is, and what it is trying to achieve, many immediately think about what technologies would help to implement it. Data virtualisation software is a good candidate it can be used as a common platform by people in different business domains.

Data virtualisation can be used to produce, consume, and govern data products

Data virtualisation can be used in:

- Connecting disparate data sources into domain-oriented development of data products
- Consumption of data products
- The implementation of federated data governance

These are discussed in more detail below.

DOMAIN ORIENTED DEVELOPMENT OF DATA PRODUCTS USING DATA VIRTUALISATION

Agile Creation of Data Products

Data virtualisation allows data products to be created using virtual views that integrate data without the need to copy it

Data virtualisation enables the agile creation of data products by allowing people in business domains to create virtual views of data in one or more data stores irrespective of whether they are on-premises, in one or more clouds or the data is in a software-as-a-service (SaaS) application.

The ability to define virtual data views quickly and easily that map to and access data in underlying data sources without copying data provides agility and simplifies access to reusable data.

Business Definition of Data Products

The first step in building data products is to define common business data names and data definitions in a business glossary so that the meaning of data shared in a data product is clearly understood and documented. Common business data names are needed for each data product. The business glossary is typically held in a data catalog. This can be within the data virtualisation server data catalog, or it could also be imported into a data virtualisation server catalog from a third-party business glossary offering.

Common business data names and data definitions for data products need to be defined in a busines glossary so the meaning of the data is clearly understood

A quick way to start the creation of a business glossary and to identify potential data products is to use a data concept model. This provides business data entities e.g., customer, supplier, orders, shipments, payments etc., that are central to your business. Data owners can then be assigned to own specific data entities. Each owner can then form a data governance working group together with subject matter experts (SMEs) who work with that data (e.g., customer data) every day, to complete the definition of all attributes that describe a data entity. Different domain-oriented data owners and working groups can work on different data entities. Once a data entity is defined in the business glossary, it then becomes possible to create core data products from which others will stem.

Creating Virtual Data Views for Data Products

Data products can be created by defining a virtual view of data entities in a data virtualisation server using the data names defined in the business glossary. The data catalog provides metadata on source data needed to create each data product. Identified data in multiple data sources can then be mapped to the schema of the virtual view describing that data product. This approach allows you to *incrementally* create a layer of **semantically linked** virtual data products that can be:

Data virtualisation allows virtual data products to be created and consumed for different analytical use cases

- Used as input to create other virtual data products
- Rapidly assembled into different virtual schema optimised for different analytical workloads
- Queried to provide data for machine learning model development in data science
- Consumed via different interfaces e.g., SQL, REST, ODATA, GraphQL

This is shown in Figure 7.

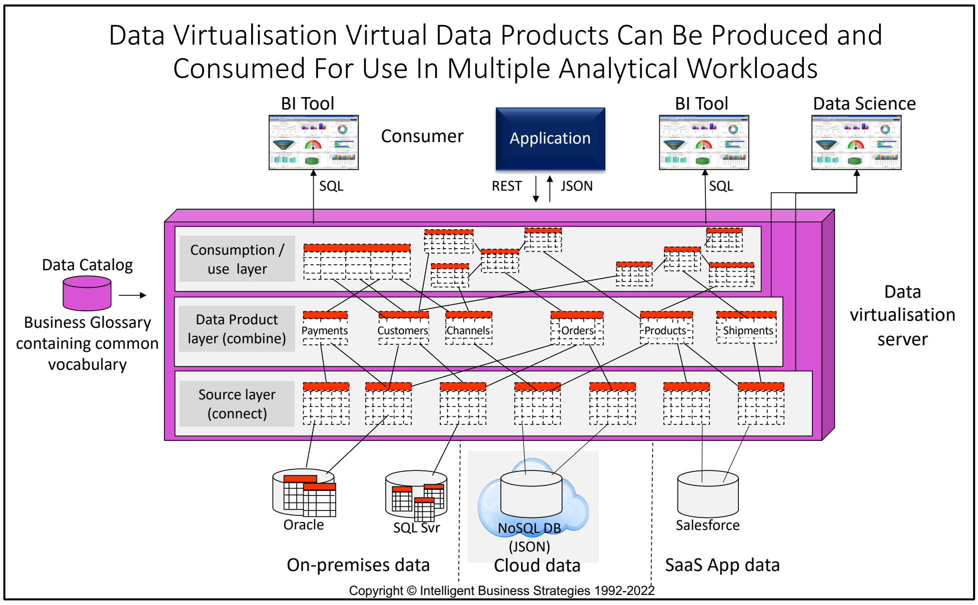

Data Virtualisation Virtual Data Products Can Be Produced and Consumed For Use In Multiple Analytical Workloads
BI Tool
Consumer
Application
BI Tool
Data Science
SQL
REST
JSON
SQL
Consumption / use layer
Data Product layer (combine)
Source layer (connect)
Payments
Customers
Channels
Orders
Products
Shipments
Data Catalog
Business Glossary containing common vocabulary
Data virtualisation server
Oracle
SQL Svr
NoSQL DB (JSON)
Salesforce
On-premises data
Cloud data
SaaS App data
Copyright © Intelligent Business Strategies 1992-2022

Figure 7

Note that this ability to quickly 'assemble' semantically linked data products to create other virtual schema signals a major step change in productivity that increases agility and shortens time to value. Data virtualisation providing business ready data to jumpstart multiple analytical projects as well as enabling component-based schema development.

Enabling Domain-Oriented Schema Evolution of Data Products Using Data Virtualisation

Data virtualisation supports data product schema evolution and versioning

In addition, data virtualisation allows domain-oriented data owners to evolve data product schema over time. This can be done by creating new virtual views of data products that could contain additional data from one or more new data sources for example. It has been possible to add new data attributes quickly and easily to virtual views in a data virtualisation server for many years. This can be accommodated in many cases without impacting anything. As a result, different versions of data products can be supported using different virtual views.

Publishing Data Products in a Data Mesh Using a Data Marketplace

If multiple domain-oriented teams across the enterprise are building trusted, commonly understood, business ready data products, it must be possible to publish these data products somewhere so that data consumers can see what business ready data is available to use and reuse in different analytical initiatives across the enterprise. To make this possible there needs to be an internal “data marketplace” - a data catalog capability – that provides access to business ready data products. It should be possible for consumers to access this data marketplace to search for and find published, business ready data products in a Data Mesh that are available for consumption (See Figure 8). In addition, information about the data owner of each data product is needed so that data consumers can request permission to access and consume data products from their respective owners.

A data marketplace provides a catalog of published business-ready data products available for consumption

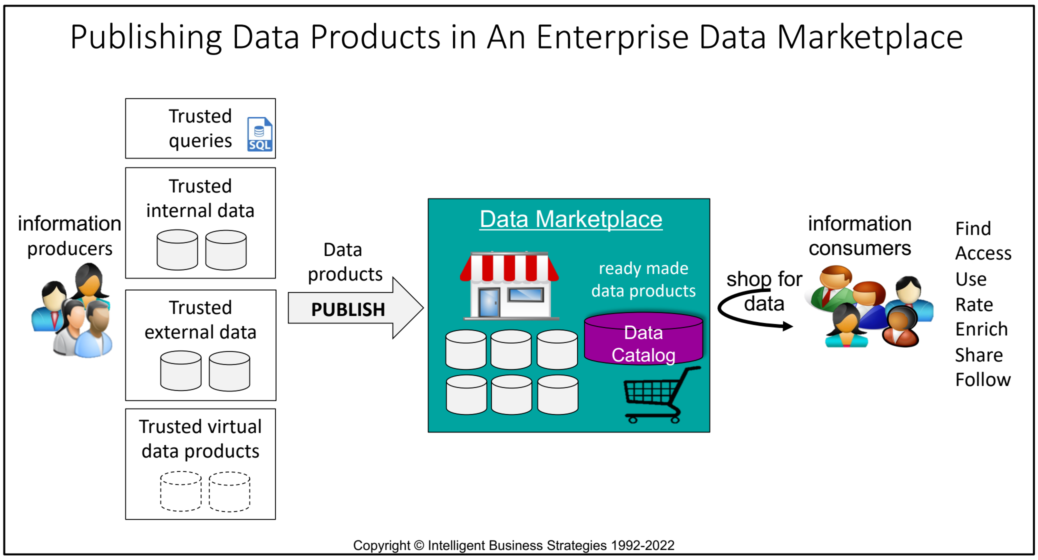


Figure 8

A data virtualisation server catalog can be used to create a data marketplace

Data virtualisation can provide this capability through its data catalog and by auditing all consumer data product access requests, all permissions granted and all queries that consume data.

A data marketplace capability is critical to success of a Data Mesh. In addition, given that multiple domain-based teams are responsible for producing their own data products, there needs to be a standard process to ensure that the publishing of data products is governed. This process would mean that each domain-based data owner quality assures data products before approving and publishing them to check that the following is available:

Data products need to be quality assured before they are published for consumption

- High quality data
- Common data names documented in a business glossary
- Full metadata lineage showing how the data product was produced
- An assigned owner of the data product
- Defined policies governing access to the data product
- Defined policies governing privacy of any personal or sensitive data
- Defined policies governing retention of data product data
- Data product version management
- A data sharing agreement so that consumers accept the terms and conditions of the agreement before being given access to the data by the data owner

CONSUMPTION OF DATA PRODUCTS USING DATA VIRTUALISATION

One of the biggest challenges with making data products available for consumption in a Data Mesh is how to avoid provisioning multiple copies of the same data which would dramatically increase data redundancy. Provisioning multiple copies of the same data adds to data complexity and could cause a significant increase in storage costs in the cloud. It also makes it more difficult to track data usage and govern that data as all copies would have to be governed in a consistent manner.

It would be much easier to manage and govern if data products were provisioned virtually since data copies would not be necessary and you could govern access to and usage of that data from one place. This is exactly what data virtualisation does.

Using A Data Virtualisation Data Catalog to Find Business Ready Data Products

By using the data catalog of a data virtualisation server as a data marketplace, data consumers can find published, business ready data products in a Data Mesh together with supporting metadata including:

Data products in a Data Mesh can be published in the data catalog of a data virtualisation server together with supporting metadata

- Common data names in the business glossary to provide understanding of the meaning of data in each data product
- Lineage to explain how data products were produced and where the data originated from
- Data product version information
- The owner of each data product
- Any data stewards associated with each data product

There may also be data quality scores, usage insights as to who is using the data product, how frequently the data is being accessed and whether or not it contains sensitive data.

Consuming Data Products in a Decentralised Data Mesh Via Data Virtualisation

Data virtualisation can provide access to physical, virtual, and stored query data products

Once a Data Mesh data product has been found, it can be provisioned virtually irrespective of whether it is a physical data product produced using other software, a virtual data product created using one or more virtual views in the data virtualisation server or a stored query that produces data for consumption when the stored query is executed. Stored queries can be saved and executed on a data virtualisation server.

The benefit of doing this is to provision data to consumers while leaving the data where it is.

Data products in a Data Mesh can be provisioned virtually to consumers to avoid the need to replicate data

Authorised consumers can access and use SQL queries, and also using web service APIs to consume Data Mesh data products. In all cases, data is provisioned virtually as opposed to creating and provisioning copies of data that have to be persisted and subsequently governed. Consumption is achieved by running SQL queries to retrieve data from a virtual data product. In addition, stored queries can be published as web services and invoked via a REST API. It is also possible to access via other interfaces e.g., GraphQL.

Consumers can be other tools such as BI servers, data science notebooks, applications, and other data integration pipelines.

AI-based smart caching accelerates query execution and consumption of data products

Some would argue that consumption of virtual data products will slow down access to data. However modern data virtualisaton servers use AI-based smart query acceleration and caching techniques to limit the impact of slow data sources and accelerate query performance.

Building New Products from Other Data Products

Virtual data products can be created quickly and can consume data from other virtual data products and data sources

Consumers in other business domains can then build new data products that can then be added to the Data Mesh as described in Figure 6. In a data virtualisation server, this can be done by combining one or more data products with data from other data sources to create a new virtual view as shown in Figure 9. Issuing queries to consume the data from the new data product, executes the steps needed to provision the required data. Querying the new data product effectively involves consumption of data from other data products in the Data Mesh.

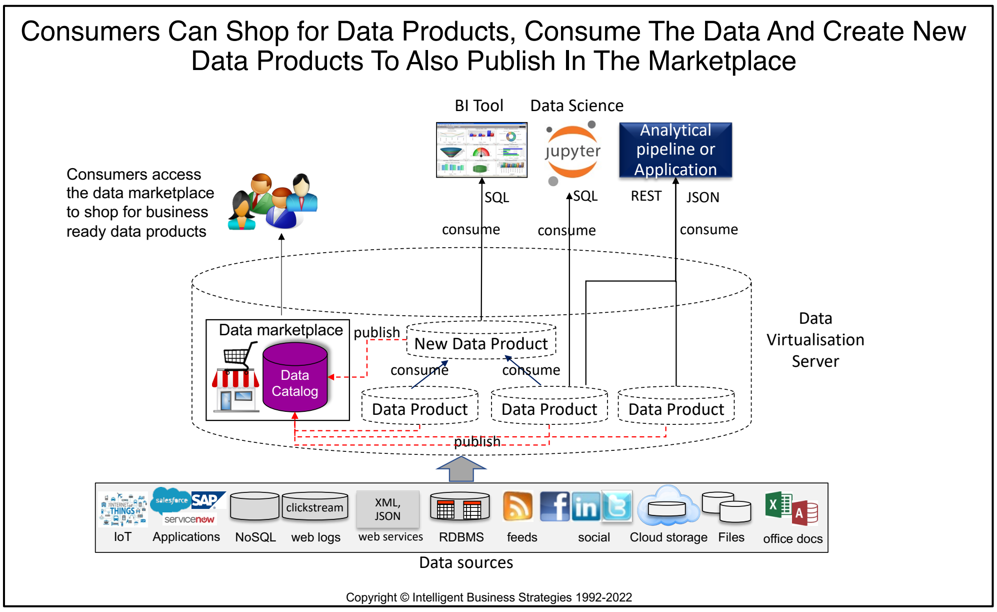

Consumers Can Shop for Data Products, Consume The Data And Create New Data Products To Also Publish In The Marketplace
BI Tool
Data Science
jupyter
Analytical pipeline or Application
Consumers access the data marketplace to shop for business ready data products
SQL
SQL
REST
JSON
consume
consume
consume
Data marketplace
Data Catalog
publish
New Data Product
consume
consume
Data Product
Data Product
Data Product
publish
Data Virtualisation Server
IoT
Applications
NoSQL
clickstream
web logs
XML, JSON
web services
RDBMS
feeds
social
Cloud storage
Files
office docs
Data sources
Copyright © Intelligent Business Strategies 1992-2022

Figure 9

IMPLEMENTING FEDERATED DATA GOVERNANCE USING DATA VIRTUALISATION

Domain based data owners are responsible for governing data products they produce in a Data Mesh

One of the key principles of Data Mesh is federated computational data governance. This is based on a data governance model that embraces decentralisation and self-sovereignty with business domains taking responsibility for setting policies to govern the data they own and the data products they produce without the need for a separate central data governance team.

Master data products are cross domain

Sensitive personal data needs to be consistently governed no matter where it is used to remain compliant with legislation

However, some data is cross-domain e.g., master data. Also, sensitive data such as personal data and financial data maybe spread across systems that are on-premises, in multiple clouds, and SaaS applications in a distributed data landscape. This data can also be used across multiple domains and requires global governance policies to apply to that data to remain compliant with regulations and legislation irrespective of domain usage. In a Data Mesh, the purpose is to produce data products that can be consumed and used widely across the enterprise. When consumed, these data products need to remain governed by policies originally set up to govern their use. However, the onus is

Domain-based teams should not be able to publish non-compliant data products in a Data Mesh

on domain-based teams to ensure that they do not publish data products for consumption that would violate compliance obligations.

A common platform is needed to govern data products in a data mesh so that policies can be consistently shared, reused, and enforced across domains

If independent domain-based teams are creating data products, then governing data via multiple independent tools is going to be a challenge. What is needed is for both global and domain-based policies to be created using a common platform so that policies can be easily shared and enforced. We need “global' data governance policies and rule specifications for data that is cross-domain (e.g., master data, highly sensitive personal data) and local data governance policies set by decentralised domain-based teams for data they own and the data products they produce.

The upshot of this is that several requirements need to be met with respect to data governance in a Data Mesh. This includes:

Global and domain-specific governance policies need to be supported

- The need for global cross-domain and local domain-specific governance policies including those associated with data access security and data privacy
- Allowing decentralised teams to inherit cross-domain policies if any cross-domain data is used in the creation of a domain-specific data product
- Allowing decentralised teams to set governance policies for data products they own and create that can be inherited by others who consume that data
- The ability to share data products without moving data
- Logging all access to and use of shared data products
- The ability to prevent personal data from being shared across borders to remain compliant with legislation that your organisation must support

Common Approach to Data Governance in a Data Mesh

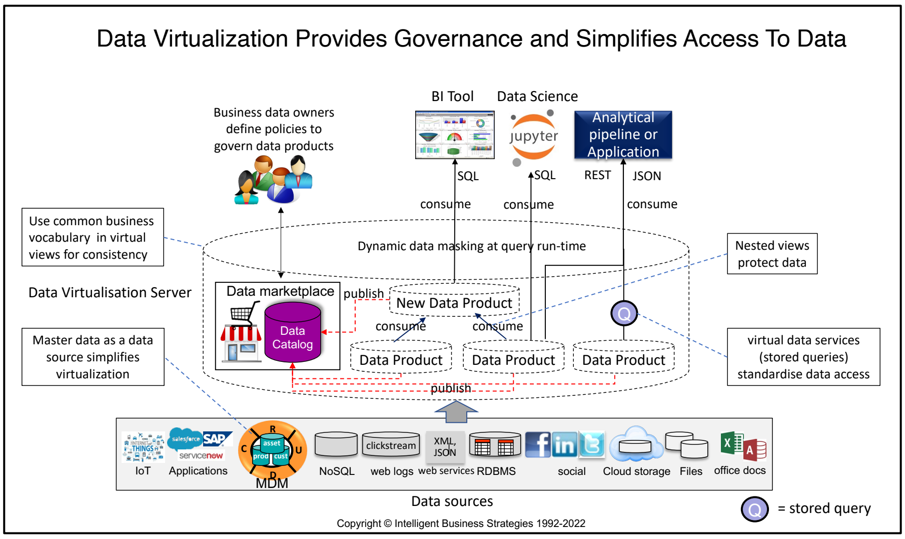
Using data virtualisation to provision data products means that governance can be controlled from one place

With respect to data governance, data virtualisation software can enable this because global and local governance policies can be controlled from a single place. Data Mesh advocates decentralised domain-based owners and teams define local data governance policies which they can all do in a data virtualisation server.

This is important because there are currently no standards to share metadata associated with data governance policies across multiple different technologies. However, if consumption of data products must be done through a common data virtualisation layer, then data can be governed from a common place. This is shown in Figure 10.

Figure 10 shows several things. The first of these is the creation of shared data services by storing queries on the data virtualization server and publishing them as web services. This has the effect of standardising data access. The second is by connecting the data virtualization server to an integrated master data management system and using it as a data source. This simplifies data integration and allow for consistent virtual provisioning of master data products across all domains. Finally, the use of nested virtual views also helps govern data because views on views of data can be used to restrict access to data to only those people authorised to see that data. This capability also provides flexibility in that multiple layers of nesting can offer up many different views of integrated data to satisfy the needs of different users and applications while protecting sensitive data at the same time.

Decentralised domain-based owners can set policies to govern the data products they own


Data Virtualization Provides Governance and Simplifies Access To Data

BI Tool
Data Science
Business data owners define policies to govern data products
jupyter
Analytical pipeline or Application
SQL
SQL
REST
JSON
consume
consume
consume
Use common business vocabulary in virtual views for consistency
Dynamic data masking at query run-time
Nested views protect data
Data Virtualisation Server
Data marketplace
Data Catalog
publish
New Data Product
consume
consume
Master data as a data source simplifies virtualization
Data Product
Data Product
Data Product
virtual data services (stored queries) standardise data access
publish
IoT
Applications
salesforce
SAP
servicenow
R
C
U
D
asset
prod
cust
MDM
NoSQL
clickstream
web logs
XML, JSON
web services
RDBMS
social
Cloud storage
Files
office docs
Data sources
Q = stored query
Copyright © Intelligent Business Strategies 1992-2022

Figure 10

Global And Local Data Governance

Master data management can be a source to data virtualisation and cross-domain policies can be defined to control virtual provisioning of master data products across the enterprise

Global and local data governance allows cross domain data products such as master data and sensitive data types to be governed by global cross-domain policies and domain-specific data products to be governed locally by policies defined by domain-oriented data owners. Master data can be a source to data virtualisation which allow you to publish physical master data products (e.g., customers, suppliers etc.) in their entirety via data virtualisation for consumption without the need to copy data and also publish subsets of master data as virtual master data products. Since this is cross-domain data, global policies can be defined in the data virtualisation server catalog to govern access to master data. Those policies can remain in place even if that data is consumed and used to create in another virtual data product.

Similarly, domain-based teams can easily create data products and define policies in the data catalog to govern access to the data products they produce and own. As a result, both global and domain-specific policies can be accommodated to govern data.

Governing Access to Data Sources

Govern access to the data sources

It is also possible to govern access to data sources via data virtualisation to ensure that data classified as sensitive or confidential can be protected within those data sources.

Sharing Data Without Moving It

Data virtualisation makes it easier to govern data sharing because data products can be provisioned without copying data

Another area that requires governance is data sharing. Data sharing is at the heart of the Data Mesh concept where data products are produced and consumed. Governance of data sharing in a Data Mesh must ensure safe, secure, compliant sharing of data with data consumers. This means that consumers need to accept the terms of use associated with a data product before it can be consumed. Also, all policies associated with the data product need to be inherited and all consumption audited so usage can be monitored.

Data virtualization makes this possible without the need to move data. By creating virtual data products, data is integrated on-demand and provisioned without copying it. Also, Data virtualisation as a solution helps multinational organisations enforce regulatory data sovereignty requirements. For example GDPR in Europe, CCPA in California or POPI in South Africa. By maintaining the data in the country/region of origin and providing governed access to the source, compliant cross-border data sharing can be accommodated.

Data Privacy

Column level access control and dynamic data masking allow data virtualisation to govern data privacy in a Data Mesh so that data products are provisioned in a compliant manner

Governing data privacy requires that sensitive personal data remains protected. In a Data Mesh, domain-oriented data owners need to ensure that this happens with sensitive personal data contained in any data products they produce. Data virtualisation software can support this using column level access control and run-time dynamic data masking to prevent personal data from being shared. This means that sensitive personal data cannot be consumed by unauthorised users and cannot appear in other data products.

Governing Access to Data Products

Data virtualisation can govern access to data products in a Data Mesh and provision data products using common business data names

Data virtualisation software supports central control of data access with respect to data sources and data products created from data in those underlying data sources irrespective of whether that data is on-premises, in the cloud or in software-as-a-service (SaaS) applications.

In addition, business users can define common data names and definitions in an enterprise business glossary within a data catalog. These common data definitions can then be used to create virtual data products in a data virtualization server. The result is that all consumers of those virtual data products will see data described using common data definitions.

CONCLUSIONS

As demand for data grows there is a need to shorten time to value by accelerating data engineering

As demand for analytics grows to improve business outcomes, so does the need to accelerate the engineering of data. To date, data engineering has been done centrally in each siloed analytical system in support of a specific analytical workload e.g., data warehouse, graph analysis, machine learning model development etc. Multiple data integration technologies are in use in each analytical system, skills are thinly spread, and the tools used are not integrated. Therefore, metadata is fractured across multiple tools and cannot be shared. In some cases such as data science, data cleansing and integration is often done using handwritten code and so in this case there is no metadata available at all. Lineage is missing. The problem with this siloed and centralised approach, is that there is point-to-point data Integration and a lot of re-invention rather than re-use. The risk of repeated, inconsistent data engineering to produce the same data across different analytical systems is very high. It is not surprising therefore that the speed of data integration pipeline development is slow, the cost is too high, and it can lead to inconsistent data. Furthermore, metadata is not easily accessible if at all. Also, the people doing data integration may not have detailed knowledge of the source data.

Current approaches are too slow, too costly, cause reinvention and can lead to inconsistent data

Data Mesh attempts to solve these problems by enabling domain-oriented teams to produce business ready data products once for consumption and reuse across multiple analytical workloads

Data Mesh attempts to solve these problems by proposing that domain-oriented owners and subject matter experts who have intimate understanding of domain data, produce and publish business-ready data products to consume and use across the enterprise. This would potentially speed up data integration pipeline development, reduce costs and make data available for reuse which would shorten time to value. However, to make this work requires that domain-oriented teams use a common platform to create data products in an agile manner.

Data virtualisation software provides and agile approach to creating data products

Data products can be governed in once place and data can be provisioned to consumers without replication

Data virtualisation software provides a way of doing this with the added benefit that data products can be provisioned without the need to create copies of data. It also allows governance of data products to be managed from a single place and policies associated with data product access security and data privacy to be shared and inherited by data product consumers. Sensitive data can be hidden by not including it in virtual data products. Alternatively, the use of dynamic data masking will prevent that data from being exposed at query runtime. In addition, new data products are quick to create, it is easy to evolve data product schema and versioning can be supported. The data virtualisation server data catalog can be used by data product consumers to find data products, understand the meaning of data in a business glossary and see lineage to understand how data products are produced. It also simplifies access and allows consumption via SQL, REST, ODATA and GraphQL all without replicating data.

Given these benefits, the data virtualisation software from Denodo is well worth considering when implementing a Data Mesh.

About Intelligent Business Strategies

Intelligent Business Strategies is an independent research, education, and consulting company whose goal is to help companies understand and exploit new developments in business intelligence, machine learning, advanced analytics, data management, big data, and enterprise business integration. Together, these technologies help an organisation become an *intelligent business*.

Author

Mike Ferguson is Managing Director of Intelligent Business Strategies Limited. As an independent IT industry analyst and consultant, he specialises in BI / analytics and data management. With over 40 years of IT experience, Mike has consulted for dozens of companies on BI/Analytics, data strategy, technology selection, enterprise architecture, and data management. Mike is also conference chairman of Big Data LDN, the fastest growing data and analytics conference in Europe. He has spoken at events all over the world and written numerous articles. Formerly he was a principal and co-founder of Codd and Date Europe Limited – the inventors of the Relational Model, a Chief Architect at Teradata on the Teradata DBMS, and European Managing Director of Database Associates. He teaches popular master classes in Data Warehouse Modernisation, Big Data, Enterprise Data Governance, Master Data Management, Creating Data Products in a Data Lake, Lakehouse or Data Mesh for Use in Analytics, Machine Learning and Advanced Analytics, Real-time Analytics, and Data Virtualisation.

INTELLIGENT
BUSINESS
STRATEGIES

Telephone: (+44)1625 520700
Internet URL: www.intelligentbusiness.biz
E-Mail: info@intelligentbusiness.biz

The Value of Data Virtualisation in a Data Mesh