Data Mesh

Data platform architecture is not a topic that is subject to the rapid changes and advances that often afflict other technology solutions. Data warehouses were first implemented in the 1980s and have been the mainstay of many data solutions ever since. We had to wait until well past 2000 for big data architectures with data lakes and parallel compute capabilities to come along. Since then, there have been some advances – e.g. in the use of real time data streaming and event driven architectures – however, there has not been a major technological shift.

One of the key features of current architectures is that they all extract data from source systems and load into a central data store – typically a data lake / data warehouse – which is the source for analytics and reporting. This centralisation of capability can lead to a number of challenges. For example:

- **Ambiguity of data ownership** – Source system data is generally owned by the team that produces the data, whereas data in a data platform is typically owned by a data platform team. The data platform team will have responsibility for data quality and accuracy in their platform however they are fully dependent on the source teams for ensuring this.
- **Bottlenecks in delivery** – With a centralised data platform, business teams can often be waiting for some time to get their requirements on to the data engineering backlog to be implemented by the data platform team.
- **Inhibiting business innovation** – Business teams that want to move quickly and develop their own analytics are often not able to when dependent on centralised analytics platforms.
- **Siloed expertise** – Data teams are expert in data integration tools but often lack the detailed understanding of the data and source systems. In large organisations it can be impossible for data engineers to be across all domains of data in any depth.
- **Management of Change** – If a source system schema is updated then it is likely that there will be an impact on the data platform, and this will often also need to be updated. However, source system owners and engineers do not always have the data platform front of mind, so it is often something that is forgotten about until too late.

The Data Mesh methodology has been proposed as a potential solution to these issues. This methodology was developed by Thoughtworks in 2019, and is based around a decentralised data platform solution with business teams taking ownership of their data product. This approach puts the power for data back with teams that own it, enabling them to move faster and be more

responsive. However, with this power comes new responsibilities in terms of making sure that data is correct and providing capabilities for others to reliably use and analyse their data.

The Data Mesh methodology is based around four key principles:

- Domain Oriented Decentralised Data Ownership and Architecture
  Business teams are fully in charge of the data domains that they are responsible for. This includes the operational creation and usage of the data as well as the analytical usage of the data. This distributes the responsibility for data to those that are closest to it and enables the business to better support continuous change and scalability. Business teams will combine subject matter expertise with data engineering capabilities to make data available to those who need it throughout the organisation. Data will be made available through defined and agreed interfaces that are compliant with global governance rules and are managed by the data domain owners.
- Data As a Product
  A key feature of the Data Mesh methodology is that analytical data that is created by each domain is treated as a product, and consumers of the data are customers. Each domain will have a data product owner that is responsible for delivery of their data product to an agreed interface and with agreed Service Level Objectives (SLOs). SLOs may include measures such as data quality, timeliness of data provided, compliance with global governance standards and ease of access. Data product owners will have a deep understanding of their data along with the data requirements of consumers. In addition to the actual data, the data product also includes metadata that describes the data such that it can be used by consumers without requiring additional information or training. The data product also includes the code and infrastructure that is used in extracting data and serving to consumers.
- Self Service Data Infrastructure
  To enable each domain to efficiently deliver data products, there needs to be a core data infrastructure implemented that abstracts the complexities of the technical implementation and enables each domain to self-serve capabilities that they need for their data products. Data infrastructure capabilities can include services such as scalable data stores, data ingestion, orchestration, data cataloguing and analytics compute capabilities. The data infrastructure also provides a baseline for security and interoperability. Infrastructure provided should be domain agnostic and should allow self-service by the business teams.
- Federated Standards and Governance
  Whilst there are many benefits to the decentralised approach to data management, it can also raise its own challenges. A major challenge is in ensuring that data domains are interoperable. To enable this, it is essential to have global standards and governance that each data domain must comply with. This enables data from disparate domains to be joined together at scale to deliver higher level insights. Domain data owners as well as data platform owners will typically comprise a governance group that will agree the standards and approaches for the platform. One of the challenges of the Data Mesh approach is determining how much autonomy each domain will have and how much governance is required to ensure interoperability. Domain teams will have flexibility to choose how they implement the defined standards and governance.

Domain data product owners and teams responsible for data products have the freedom to select the technologies that are most appropriate to their data product and the requirements of those that need it. This may include using different tools or cloud service providers to other domains in the organisation. However, all technologies must comply with the global governance and standards for the platform. And just because teams have the freedom to choose technologies doesn't mean that they must. Where common data infrastructure is provided that meets the business needs, this should be utilised as a preference to avoid each team duplicating effort by developing their own solution. This may mean that organisations still have a single data lake and data warehouse. The difference however is that these now have decentralised ownership.

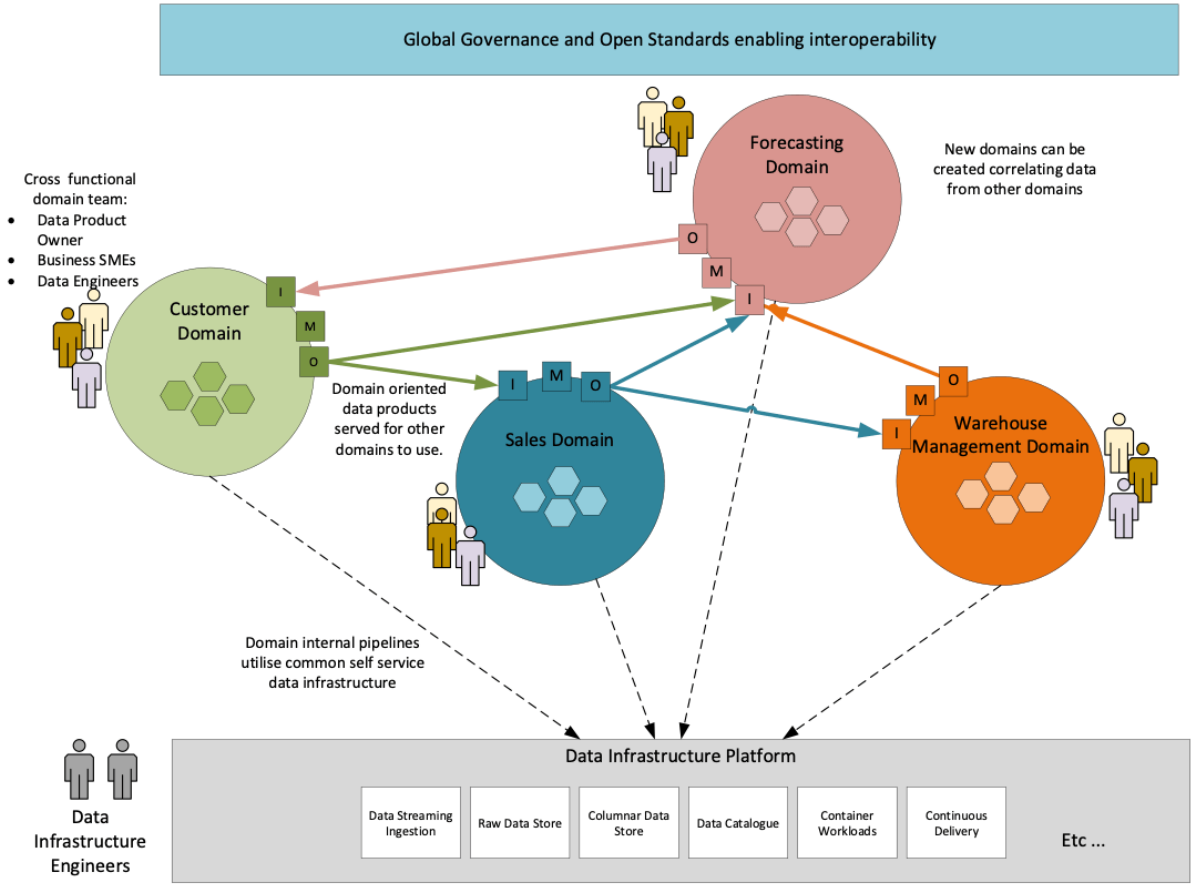
It is also important to note that not all business teams that own data products will be creators of source data. Large organisations often have analytics teams that utilise data from other domains and create their own data products. These teams will become their own domain and will have responsibilities for their data product in the same way as other domains.

Data products which are produced by each domain need to meet a set of globally agreed criteria defined by the organisation. Data must be:

- **Discoverable** – It must be easy for business teams to discover data products that are available, who owns it, where it comes from. This may be via a central data catalogue, wiki or other service.
- **Addressable** – Data must follow agreed conventions for how it can be addressed and accessed.
- **Trustworthy** – Data that is available must be reliable. Data owners need to provide a service level objective around how closely data provided reflects what actually happened. Data cleansing and integrity testing can be performed at source to ensure an acceptable level of quality.
- **Self-describing** – The data product should contain sufficient information that analysts are able to use it without requiring additional information or expertise. This can be by providing data schemas or sample datasets as well as detailed documentation.
- **Interoperable** – Data from domains across the organisation needs to be able to be joined in a consistent and reliable manner. Challenges can arise with different identifiers being used in different data product, different granularities of data being stored, and different data types being used. Global governance is required to ensure that this is managed effectively.
- **Secure** – Data that is made available must be securely managed. This may require fine-grained access control to restrict access to specific datasets, data rows or data items.

The interfaces are responsible for ensuring data quality, security and privacy enabling the consumer to focus on use of the data.

The following diagram of an example retail organisation illustrates how the various components of data mesh architecture interact.



This example comprises of four domains. Three of the domains – Customer, Sales and Warehouse Management – are business teams that own and manage source data. The fourth domain – Forecasting – performs analytics on the data provided by other teams to forecast future sales and makes that data available to them. Each domain provides Outputs (O), and Metadata (M) about the data provided. Domains may also take inputs from other domains (I). All domains make use of common infrastructure that provides ingestion processes, storage, compute resources and data catalogues. Each domain must also comply with global governance standards defined by the organisation. Each domain is comprised of a cross-functional team that includes the data product owner, business SMEs and data engineers.

Data Mesh is a lot more than a new data architecture. It is a new organisational paradigm that can enable organisations to become fully data driven operations. There are many benefits that it can bring to organisations, such as enabling business change at speed, providing a better understanding of data value and innovation opportunities, reducing complexity, democratising data creation and access and improving data quality.

However, the implementation of Data Mesh will require a major shift in mindsets as well as a cultural buy-in for the methodology across the organisation. There will also be many governance and technical challenges to be solved to implement the methodology effectively. It is certainly not an approach that will benefit every organisation. Larger organisations that need to innovate and change rapidly and have many teams creating and utilising data are likely to be the major beneficiaries of this approach.

Data Mesh is still a new concept. It will be interesting to see how it evolves in the coming years and how it gets implemented in practice. For many organisations, their existing centralised data platform will continue to deliver value to the business. Where organisations do struggle with the need to be responsive and adaptable to change using the current models, these challenges can often be mitigated at least partially through existing solutions such as effective data governance, streamlined work processes and better management of current data platforms.

If your organisation is struggling with the need to be responsive and adaptable to change using the current models, or you would like to find about more about Data Mesh, please contact a Fujitsu Data & AI specialist now.

This article has aimed to provide a high-level overview of the Data Mesh methodology and concepts. If you would like to read more about it then the following articles are recommended.

https://martinfowler.com/articles/data-monolith-to-mesh.html

https://martinfowler.com/articles/data-mesh-principles.html

https://databricks.com/session_na21/architects-open-source-guide-for-a-data-mesh-architecture

https://towardsdatascience.com/decoding-the-data-mesh-594b44b4b26f

https://www.jamesserra.com/archive/2021/02/data-mesh/

Contact

Fujitsu Data & AI

+61 3 9924 3000