Data Mesh for Trusted Public Sector Data Sharing in Singapore

/thoughtworks

aws partner network
Advanced
Consulting
Partner

Table of contents

Executive summary

Singapore is one of the few smart nations in the world, placing digital transformation at the centre of every Government function. Specifically, data has been recognized by the Government as the fuel for digital transformation[1].

Data sharing is key to becoming a data-driven nation. Without a trustable and resilient data sharing framework, it's not possible to bootstrap new initiatives and policies quickly. Delay in data access requests leads to loss of valuable government resources. Lack of transparency and controls during sharing causes cracks in data privacy and security. Last but not least, a good data-sharing framework becomes a critical factor in engaging citizens and private entities to contribute effectively to the nation's growth. The Government Data Office acknowledges this and has devised a robust data strategy to facilitate trusted data sharing.

In this white paper, we would like to propose the Data Mesh paradigm, developed by Thoughtworks, to aid the public sector in this journey to build a trusted data sharing framework, based on challenges that we've seen. We will leave you with a scalable domain-driven approach to data sharing that enables public sector agencies to share transparent and trustworthy data with ease. This in turn, will help create a network of data driven agencies that can collaborate to create a strong data sharing ecosystem at the national level.

[1] Daniel Lim Yew Ma (8 Aug 2019), Bringing Data into the Heart of Digital Government

Data sharing in Singapore public sector

Singapore Government Data Strategy

The Government Data Office was set up under the Smart Nation and Digital Government Office (SNDGO) to execute the Government Data Strategy (GDS) by 2023. The primary goal of the GDS is to enable discoverability and accessibility of core government data assets within 7 days of a data request, thus facilitating:

- Data driven policy making
- AI/ML solutions for public good

In order to facilitate the sharing of trusted data assets across agencies, the GDS aims to build the right infrastructure for data sharing by building Trusted Centres (TC) to be data intermediaries for individual, business, geospatial and sensor data. These TCs serve as aggregators of datasets from various public sector agencies (referred to as SSOT or Single Source of Truth, for a given domain). For future references in this white paper, we'll be using SSOT to refer to the agency and TC to refer to the Trusted Centre, quite extensively.

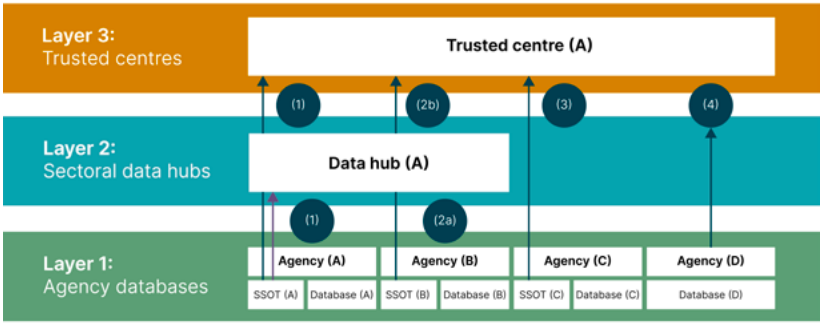

Fig 1a. Overview of the new Government Data Architecture[1]
Source: Smart Nation and Digital Government Office

[1] Daniel Lim Yew Ma (8 Aug 2019), Bringing Data into the Heart of Digital Government

A comprehensive digital infrastructure is being set up to support this framework. Some of the offerings include:

- Whole-Of-Government Analytics (WOGAA) platform that enables the rapid development of data and AI models
- Vault.Gov.SG as a data discovery and exploration platform

At the core, the data sharing strategy aims to be agile to support new use cases in a nimble fashion. We recognize that GDS has placed great importance on decentralized and federated approaches for data sharing. This white paper aims to emphasize and enhance these aspects through the use of Data Mesh.

Data Mesh paradigm to enable trusted data sharing

On public sector data sharing, the Trusted Data Sharing Framework[2] and the Digital Government Blueprint[3] outline the key objectives and principles of data sharing as below:

Fig 1b. Objectives and Principles of trusted data sharing from Trusted Data Sharing Framework[2] and Digital Government Blueprint[3]

[2] IMDA (15 Oct 2020), Trusted Data Sharing Framework
[3] GOVTECH SINGAPORE (Dec 2020), Digital Government Blueprint

The key objectives of data sharing being to improve culture of excellence around sharing and using data securely, and raise public officers' competencies in safeguarding data; enhance frameworks and processes to improve accountability and transparency of the public sector data security regime and to introduce and strengthen organisational and governance structures to drive a resilient public sector data security regime that can meet future needs.

Data Mesh, explained in detail in the next section, is a decentralized socio-technical approach in managing and accessing analytical data at scale. In other words, Data Mesh recognizes the inherent challenges with interactions between people and technical organizational systems. In this white paper, we intend to address how the following objectives derived from the ones above, could be achieved using 'Data Mesh' as a paradigm:

- Enabling domain-oriented accountability in data sharing
- Enabling source agencies to share richer and interoperable datasets
- Establishing comprehensive data lineage and quality to improve transparency and trustworthiness during sharing
- Strengthening organizational and governance structures to build resilient sharing frameworks
- Example reference architecture from AWS to support the above

In the next section, we'll understand what Data Mesh is, what issues it intends to solve and the principles of this paradigm.

Data Mesh - an Introduction

In this section, we go through excerpts from the Data Mesh literature[4, 5] on what the paradigm is about and how it's various pillars come together to drive business value, at scale.

The big data dysfunction

According to a NewVantage Partners Study[6], big data continues to be a struggle for most enterprises. The survey reports that only 24% of firms claim to have succeeded in creating a data-driven organization. Only a similarly small minority of companies claim to have successfully built a data culture within their organizations. Yet technology is not why they are failing. Over 90% cite people and processes as what stands between them and transforming their organizations through data.

Traditional approaches to generating big data insights have resulted in unhappy stakeholders. Some of the reasons include:

- Data quality issues
- Not factoring in people and processes
- Problems with sourcing data engineering talent
- Overworked data platform teams, and
- Endless and costly data infrastructure projects

[4] Thoughtworks (2021), Whitepaper: The Data Mesh Shift
[5] Zhamak Dehgani (20 May 2019), How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh
[6] NewVantagePartners (2021), The Journey to Becoming Data-Driven: A Progress Report on the State of Corporate Data Initiatives

Consequently, organizations exhibit a common set of failure modes.

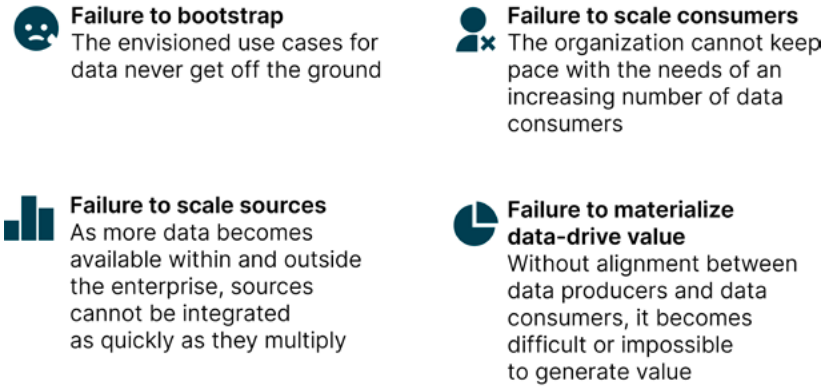

Fig 2a. Failure modes exhibited by many organizations owing to big data dysfunction

A big silo and an insurmountable bottleneck

The current data architecture paradigms, namely the data warehouses and data lakes, are designed to enable access to all the enterprise' data. However, they tend to create silos not only owing to technical constraints but also due to organizational ones. Namely, the teams that build and operate these monolithic repositories must be populated with specialized data and infrastructure engineers and are often divided based on skill rather than business outcome. These engineers will be tasked with getting data from people and teams across the organization that have little incentive to ensure they are sharing only correct, trustworthy, and meaningful data.

The current data architecture paradigms also fail beyond a certain pivot point in terms of team size, complexity of data, tools and processes.

A paradigm shift that goes beyond technology

To remedy this situation, one must go beyond a system-oriented paradigm towards a domain-driven one.

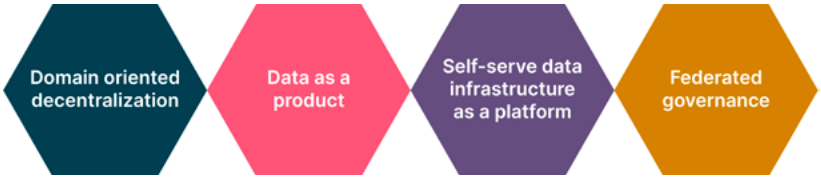

Fig 2b. Pillars of the Data Mesh Paradigm Shift

Transitioning to domain-driven data products

Data Mesh introduces the concept of a data product (DP) as its smallest unit of architecture that can be independently deployed with high functional cohesion, and includes all the structural elements required for its function.

The paradigm calls for thinking of data as an asset and an end-to-end use case or a product. Compared to other prevalent paradigms such as data lake or data warehouse, the responsibility and accountability of cleansing the data and making it fit for purpose, all lie closer to the producers of the data.

For instance, in a Health promotion agency, health activity data needs domain knowledge to ensure it can be processed effectively and made ready for analysis. A central platform team will not have such knowledge and requires the health activity team to own the process of making the data fit for consumption.

A data product should be as consumer-driven as possible to help serve the wider organization, including data analysts and anyone else who needs to work with it.

The architecture and teams are decomposed based on domains and could be source aligned (representing the raw data at the time of creation), consumer aligned (fitted or modeled for a particular consumer) or shared.

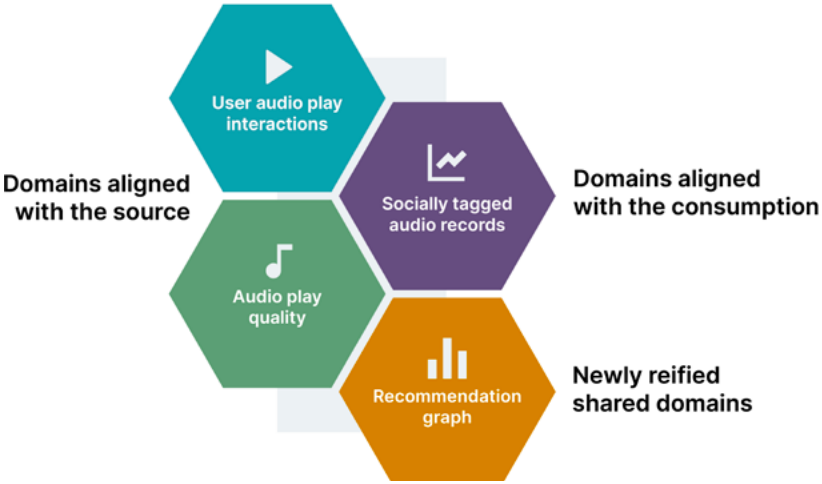

Fig.2c Decomposing the architecture and teams owning the data based on domains - source, consumer and newly created shared domains

For such a decentralized, domain-oriented approach to succeed, several prerequisites must be met. A domain can have a variety of data products but the characteristics of any domain data product remains intact. The characteristics of a domain data product are:

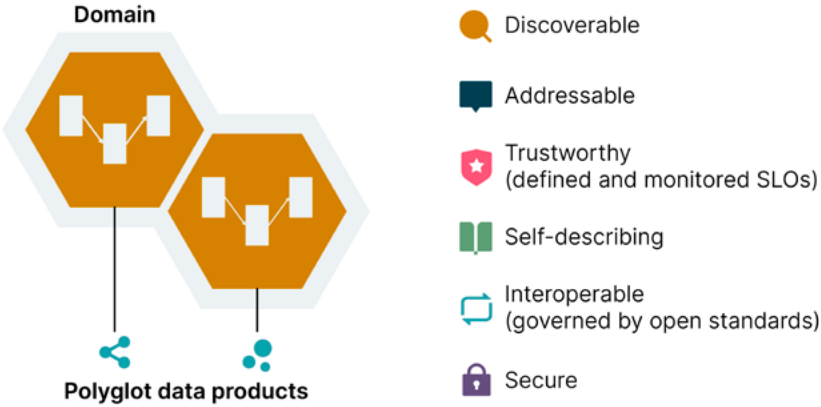

Fig.2d Characteristics of domain datasets as product

Once achieved, these qualities allow the model to scale rapidly with independent yet cohesive data products.

Self-service data infrastructure as a platform

In an effort to enable each domain to easily build data products and also tap into existing data lake infrastructure setups, Data Mesh offers data infrastructure as a service. Instead of requiring each domain team to engineer its own data platform, the necessary domain-agnostic base infrastructure is provisioned from a self-service platform, while allowing for domain-specific customization. This gives the teams a high degree of autonomy while also allowing the integration of central assets such as an existing data catalog.

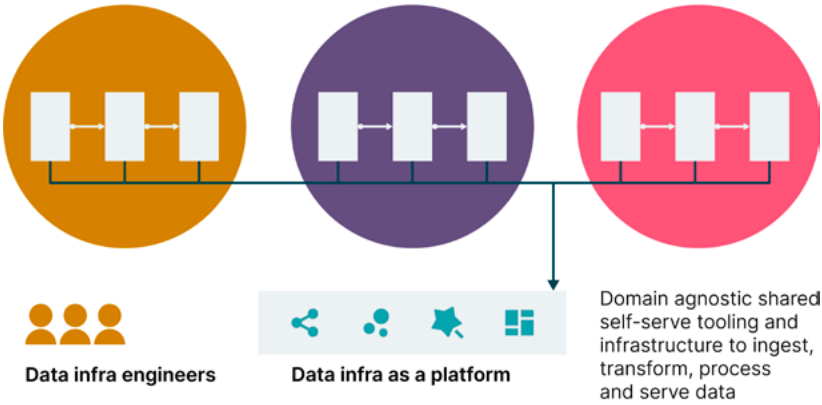

Fig.2e Extracting and harvesting domain agnostic data pipeline infrastructure and tooling into a separate data infrastructure as a platform

Federated governance

One of the core features of the Data Mesh is its federated governance model that achieves interoperability through standardization. A supportive organizational structure, incentive model and architecture is necessary for the federated governance model to function: to arrive at global decisions and standards for interoperability, while respecting autonomy of local domains and implement global policies effectively. Only with interoperable data can data analysis involving multiple data products lead to valuable insights and action. The actions can influence the formation of the next cycle of data, establishing a connected cycle of intelligence.

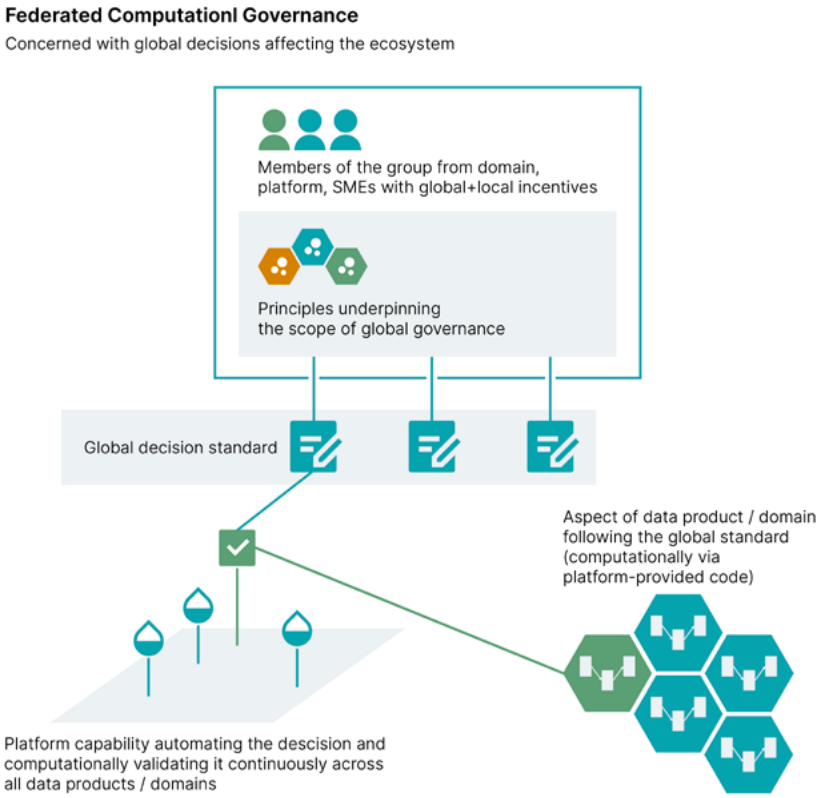

Fig.2f Federated Computational Governance Model

Complementary nature of the Data Mesh pillars

The four pillars of Data Mesh (Fig. 2b) complement each other to bring together this paradigm.

The self-service data infrastructure platform enables creation of data products at ease.

The federated governance principles help standardize the domains and their data products, yet provide domain autonomy. These principles are baked into the self-serve platform to ensure

that the data products adhere to the set guidelines. The domain data products and their feature sets push the boundaries of the capabilities that need to be served by the platform and contribute to the overall governance policies. This complementary nature of the pillars come together to create a feedback-driven evolution that makes Data Mesh the sum of its capabilities. This helps facilitate creation of faster insights and speeds up innovation.

In the upcoming section, we will dive into how Data Mesh principles can be applied to solve the objectives laid out for trusted data sharing in the Singapore public sector earlier in this paper.

Data Mesh for enabling trusted data sharing in the Singapore public sector

Enabling domain oriented accountability in data sharing
The Data Mesh approach advocates a domain-oriented decomposition and ownership of data systems over activity-oriented teams with technically partitioned architecture as is the case with most data platforms today.

A domain-oriented approach allows teams to take ownership of their own deliverables by reducing dependencies and organizational silos through practices that facilitate collaboration and open knowledge sharing. Having accountability and ownership of data assets is key to the success of a data platform to ensure trustworthy and compliant data.

How do domains help to create accountability and how to carve them out?

Let's consider the scenario of multiple SSOTs, publishing route data for bus, train and others in the transport sector.

As part of the Data Mesh paradigm, each of these SSOTs could be considered as individual domains in a sector. These will be operated by individuals who understand the domain very well and have the right set of tools and frameworks to align closer with the different purposes being served by them. This will enable the SSOTs to bootstrap new use cases and scale rapidly instead of depending on Whole-of-Government (WOG) teams to enable their use cases.

A domain may have one or many data products falling into the following categories:

- Source aligned (representing raw data at the time of creation, harmonized and fit for consumption)
- Aggregated (curated data across multiple data products and domains)
- Customer aligned (derived data, targeting very specific customer needs)

In the case of the Singapore Public sector, a transport Trusted Centre (TC) can collate data from two or more different SSOTs and enable it for consumption by public sector officers. This becomes an aggregated domain data product.

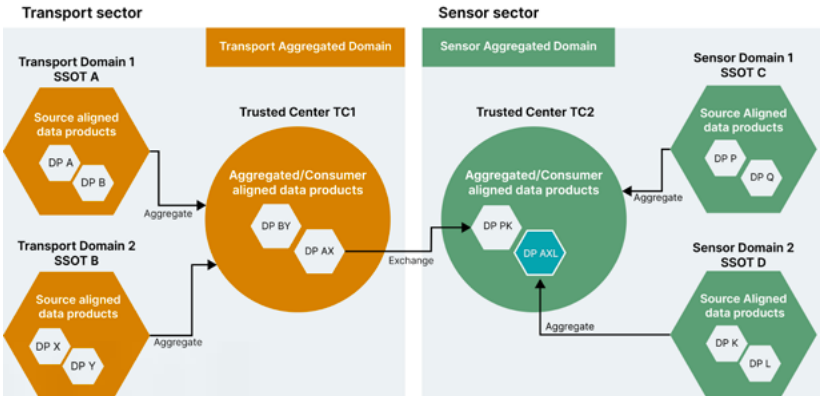

Fig 4a. Example of domain-oriented categorization of SSOTs and TCs in the Singapore Public Sector

In order to establish accountability and ownership within these domains, it's important to set the appropriate roles with their respective responsibilities. The key roles (non-exhaustive) that Data Mesh suggests within these domains are:

- **Data Owners** are responsible for granting access and

defining policies for their respective domains. In a public sector context, the Data Owners of a TC work with the Data Owners of the SSOTs to formulate governance policies and standards specific to the given sector and enable execution of the same by communicating to the data domain stewards, data product owners and the self-serve platform teams.

- **Data Domain Stewards** help monitor the quality of the data products and ensure there is transparency via rich lineage provided by the products. In the example given above, the data domain stewards of the Trusted Centre would work with data domain stewards of the SSOTs to enforce data quality and lineage standards using the common tools and APIs provided by the self-serve platform.
- **Data Product Owners** for each data product (in a SSOT) collaborate with other product owners to advance the features of the product. The Product Owner is also responsible for quick remediation of issues reported by users on a given data product. They're vital in carving out the right features and making this product a delight for the users to consume.
- **Data Product Developers** are concerned with engineering the internals and implementation of the data product.

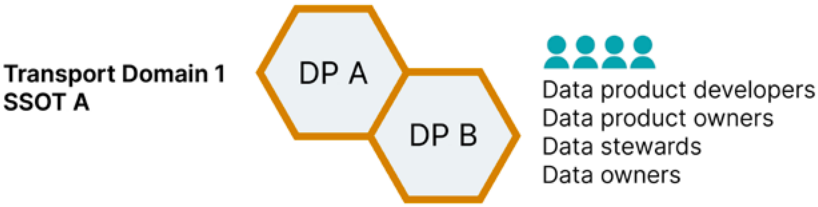

Fig 4b. Roles associated with a domain and it's products

Each of the above roles may be shared by multiple individuals, depending on the complexity of the domain data products, ensuring that each action within the Data Mesh sharing framework is accountable.

Enabling source agencies to share richer and interoperable datasets

In the previous section, we've created domains in the transport sector. Here, we move on to create rich data products within this domain. As we saw earlier, a good data product should be discoverable, addressable, accessible, interoperable, secure, transparent and trustworthy.

The following sections detail how we can achieve the creation of such a data product, in a nimble way.

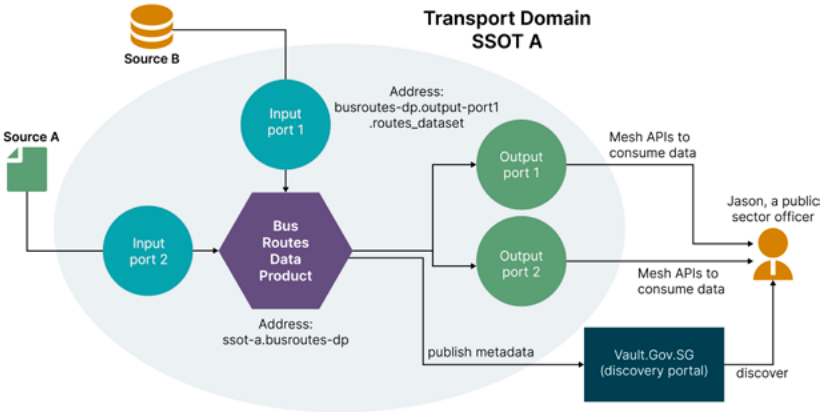

Fig 4c. Example of a data product in the transport domain

Discoverability refers to the ability to easily search and find the required datasets by a consumer. The 'bus routes' data product can be enabled for discovery by pushing the relevant information such as metadata, lineage, ownership information, freshness of data et al to the central data catalog (Vault.Gov.SG).

The self-serve platform captures this information in a consistent manner across domain data products by setting metadata standards to enable discovery and making them available via the data catalog.

Making use of both these paradigms, Jason, a downstream consumer can now discover the data product 'ssot-a.busroutes-dp' and understand if the data is fit for their use case. If deemed fit, Jason can send a data request via the product API to the data domain steward, Grace, a public sector officer.

Addressability refers to the ability to uniquely identify and access a data product.

Let's consider the 'bus routes' data product as shown in the figure above. An input port in this context may be a source system outside of the product or the output of an upstream data product. An output port similarly stores the data produced by a data product using a suitable underlying storage.

In order for the data product to be addressable, the primary step would be for the domain to uniquely assign a name reflective of the purpose of the product. For instance, 'busroutes-dp' can refer to all the datasets within that product using notions of relatedness. Additionally, if we're serving the same datasets on different output ports, we can qualify the output port as well in the name, for instance 'busroutes-dp.output-port1.routes_dataset' illustrated in Figure 4c points to a dataset on output-port1. Prior to publishing to the data catalog, the unique name of the domain will be prefixed to the product name such as 'ssot-a.busroutes-dp.routes_dataset' so that this product and dataset is uniquely addressable across the public sector.

Accessibility is the ability for an end user to raise data requests with ease and have authorized access of data, suitable for consumption.

Now that Jason has raised the request to consume 'ssot-a.busroutes-dp', Grace receives the request through an

automated governance workflow, informing her of the purpose of usage and actions on the request. Upon approval, Jason is automatically granted access to the datasets in the request and is now authorized for data consumption as illustrated in Fig 4c. In this case, Jason could connect to the relevant output ports to consume the data.

Interoperability is the ability for various data products to create, consume and exchange data with each other with clear and shared expectations.

Now that Jason has access to 'ssot-a.busroutes-dp', he can consume the data from relevant output ports and link this data with a few more data products to arrive at aggregated insights.

With many domain entities, it's important to have standardized business definitions aka common business polysemes, established between data owners of different data products (SSOTs). The Trusted Centre should be responsible for setting up these common business definitions across domains in a given sector. The TC data domain steward's job is to identify the common polysemes and relationships that need to be curated by data products to make it easy to join data across different domains.

Security is the process of protecting data products from unauthorized access and corruption. Here, we'll touch upon how data classification and auditability can be achieved with the Data Mesh paradigm.

Data classification is an integral part of defining a secure data product. Going back to our example data product, 'ssot-a.busroutes-dp', Grace, the data domain steward, is responsible for classifying the data fields according to WOG standards

set in IM8[7] and PDPA[8] and the domain standards set by the respective TC. For instance, Grace may define a bus staff dataset within this product, to have columns with Personally Identifiable Information(PII) data. Based on this definition, the self-serve data infra platform ensures that the right privacy policies are applied on the output ports of the data product prior to end user consumption.

Audit and Compliance are essential to serve as security guard-rails of a data product.
The data product 'ssot-a.busroutes-dp' captures all the transactions, policies applied on the PII datasets and authorization logs into a central storage that is accessible by the data product but managed by the self-serve platform. The TC security officers can use this data to generate audit and compliance reports for various data products across the data domains (SSOTs).

In addition, through data lineage and the notion of all data products being discoverable and transparent, the data platform can identify and flag datasets that do not have, for instance, a PII policy yet consume PII data exposed by an upstream data product.

Empowering data product teams

In order to build good data products, it's important to facilitate and empower the data domain product teams to have complete control of all aspects necessary to deliver their business goals. This is often restricted by having partitioned teams according to technical responsibilities.

To enable domain independence, the Data Mesh paradigm advocates using self-service infrastructure and providing the

[7] Singapore Government Developer Portal, Instruction Manual (IM8)
[8] Personal Data Protection Commission, Personal Data Protection Act

data product developers with complete control of their own data and infrastructure. The necessary guard rails and helper libraries are provided by the platform. This allows the data product teams to build features end-to-end at a pace they are comfortable with, with minimal reliance on external teams.

Establishing comprehensive data lineage and quality to improve transparency and trustworthiness during sharing

Let's say Grace, the public sector officer and data domain steward in the previous example, needs business statistics from a few datasets to arrive at a critical decision. She finds that the stats don't tally with what she was expecting. She wants to find out where this data comes from and what processes have run on it to ascertain if the data is trustworthy. However, she is unable to find the relevant information and consequently is unable to make the decision on time.

Data lineage to ensure transparency during data sharing

Data lineage is the information that helps users to understand the entire journey of data from source to destination with the history of changes that have taken place in between.

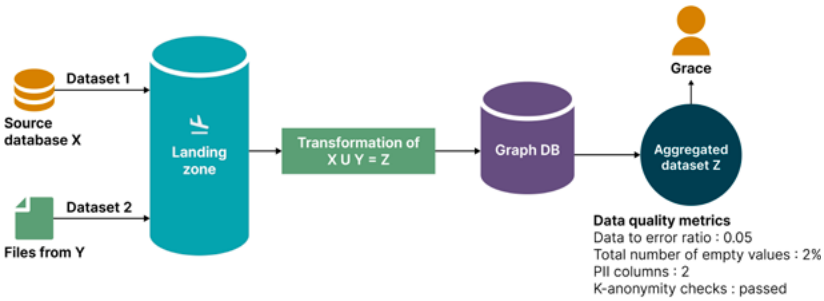

Fig 4d. Example of a data product in the transport domain

In Grace's case, a data lineage indicated as above would enable her to understand where the data comes from and what transformations have been done on the datasets to provide the aggregated dataset Z. This helps her to assess and remediate the trustworthiness of the resultant dataset.

Enforcing lineage with the self-serve data platform

The self-serve data platform helps to enforce cross functional concerns such as lineage on the datasets produced by domain data products. The input ports, the transformation framework and the output ports are served by the data platform as capabilities that the domain data products must use to capture lineage of data across the board. However, it's the data domain steward's (SSOT's) responsibility to make sure that their respective domain data products are exposing comprehensive data lineage to their users.

Despite the federated nature, the platform has the ability to bake in the necessary controls into the system. For instance, a dataset that does not publish lineage information back to the platform can have its data flagged as unfit for consumption and be notified for it. The TCs can also then play the role of the intermediary with automated audits and reports to notify the respective SSOTs of these issues in the system.

Enforcing federated data quality

Even when the data journey is clear, it is not possible to trust the data if the data quality does not meet the appropriate standards. With the quality metrics stamped as shown in the above figure, Grace can determine whether this dataset is trustable or not and call for remedial action on the required stages to fix the data quality issues.

Data quality predominantly includes structural checks, schema checks and domain specific business rules.
In a federated fashion, these can be facilitated at the self-serve platform level but governed at the domain data product level.

Structural and schematic checks

In the self-service platform, when a data product input port is created, necessary structural checks can be performed on the incoming data based on the ingestion configuration registered during the data product creation time.

Similarly, based on the metadata registered at the time of creation of the data product, schematic validations can be performed at the data product input port, by the self-serve data platform.

Domain specific business validations

Each domain data product must take care of validating their incoming and outgoing data based on domain specific business rules. The TCs are best placed to curate and set the quality benchmarks to be met by the respective SSOTs for a given domain. For instance, the Transport TC is responsible for charting out the quality metrics needed for the transportation datasets and each Transport SSOT data domain steward needs to ensure that their datasets meet the quality standards set by the respective TC.

In parallel, there will also be a set of WOG standards such as PDPA guidelines and IM8 that will need to be adhered to by each data product.

These standards can be consistently enforced by the self-serve platform across all domain data products.

Incentivizing domain data stewards to provide data products of rich data lineage and quality

A comprehensive data lineage and good data quality will naturally increase the trustworthiness of data in SSOTs and TCs. However, it is necessary for the overall sharing framework to come up with incentives to motivate the data owners, stewards and the product owners to provide datasets of high quality and rich lineage. These incentives need to be defined closer to a sector. For instance, a periodic audit of the quality of the datasets should be done by the TCs on the SSOTs to ensure that they're following all the benchmarks. Rewards can be given out as public announcements if the baselines are met. As an alternative measure, penalties can also be introduced if the domains are not meeting the necessary measures.

Strengthen organizational and governance structures to build resilient sharing frameworks

Federated governance

The Data Mesh paradigm emphasizes federated governance balanced between global and domain governance.

In the context of the Singapore Public Sector, the global standards would be set by the Government Data Office (GDO) for all SSOTs and TCs. The TCs and SSOTs manage the governance in each of their domains.

The enforcement of the policies where possible should be implemented by the self-serve platform to ensure that the standards set by the GDO can be unequivocally enforced for all domain data products. These policies should evolve over time based on feedback and legal compliance changes. Sector specific data governance policies should be curated by the TCs as mentioned in the previous section and enforced by each SSOT without deviation.

The following diagram represents an example of balancing the federated governance:

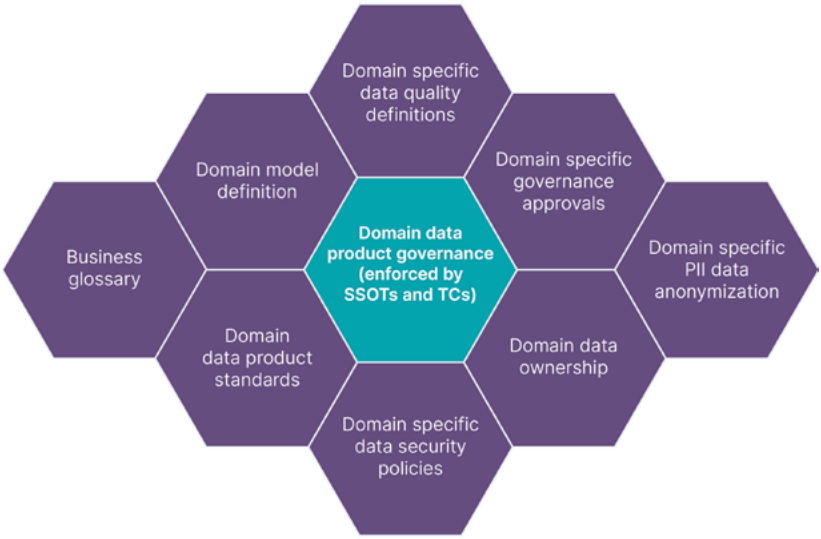

Fig 4e. Exemplar capabilities falling into domain and global governance

Operating Model for Federated governance

The federated governance approach enables the following advantages:

- Ability to focus on specific data entities, agency challenges and sector priorities
- An easier model to implement initially and sustain over time
- Resolution of issues at the SSOT level without involving the GDO at all times
- Each SSOT and TC empowered with decision making
- Breaking information silos and enabling cross sector linkages, while still retaining ownership with SSOTs
- Enforcement of global compliance including privacy policies by the GDO

Autonomy at the domain level can be coordinated by regular data governance working group check-ins. These check-ins can bring together the SSOTs, TCs and the GDO on the same page on evolving policies. The working group should also prioritize the automation of new policies as and when they evolve, on the self-serve data platform.

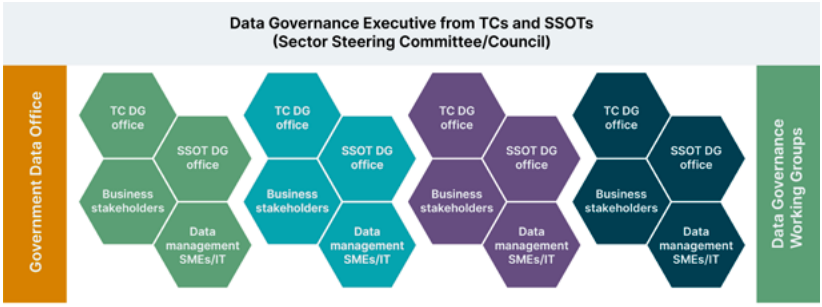

Fig 4f. Exemplar operating model for federated governance

Implementing governance policies on the self-serve platform

Governance is a combination of people, process and technology. It's important for the people to enforce the right processes and wherever possible implement them through technology. All the concerns that can be automated should be served out as features, controls and architectural fitness functions[9] (an architectural fitness function provides an objective integrity assessment of some architectural characteristics, which may encompass existing verification criteria, such as unit testing, metrics, monitors et al) on the self-serve platform.

For instance, the self-serve data platform should provide the following capabilities to support federated governance, to name a few:

- Implement the access control restrictions on data products via authentication, authorization and privacy controls
- Manage a data catalog across all domain data products
- Provide data lineage across domains
- Facilitate cross domain data requests via workflows
- Enable privacy preserving data linkage across domains

[9] Thoughtworks (May 2018), Technology Radar, Architectural Fitness Function

The data platform capabilities pertaining to governance can be classified into different planes as below.

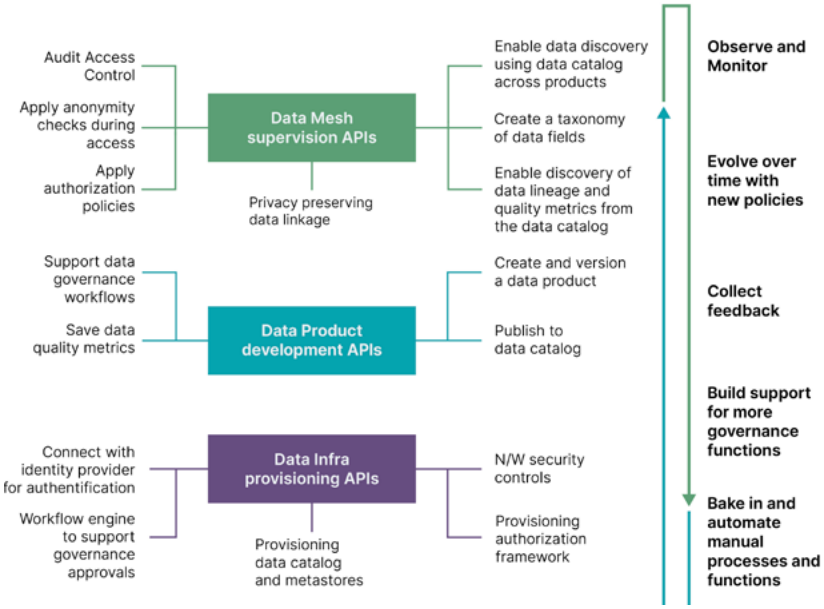

Fig 4g. Exemplar federated governance capabilities that could be served from the various planes of the self-serve platform

The self-serve platform enables the ability to bake in the defined governance procedures and policies and enforces them at various stages of a domain data product since its inception.

With the updates to existing policies, the self-serve platform also needs to evolve and the updates to be automatically pushed and enforced on all the data products.

Summarizing the conceptual view

In this section, we've covered how the various objectives set out earlier in this white paper could be envisioned through the Data Mesh paradigm.

For successful data sharing via Data Mesh, all the four pillars need to come together as one:

- Enabling domain-driven decentralization through federated roles within SSOTs and TCs
- Building rich domain data products following the shift-left paradigm, with the ownership of serving the data assets, end-to-end residing with the data product teams
- Self-service data platform that provides the APIs and services to build rich data products and evolves and scales over time with the customer needs and governance policies
- A federated governance with domain governance shared by TCs and SSOTs and WOG governance spearheaded by the GDO with the policies evolving over time to support the compliance landscape

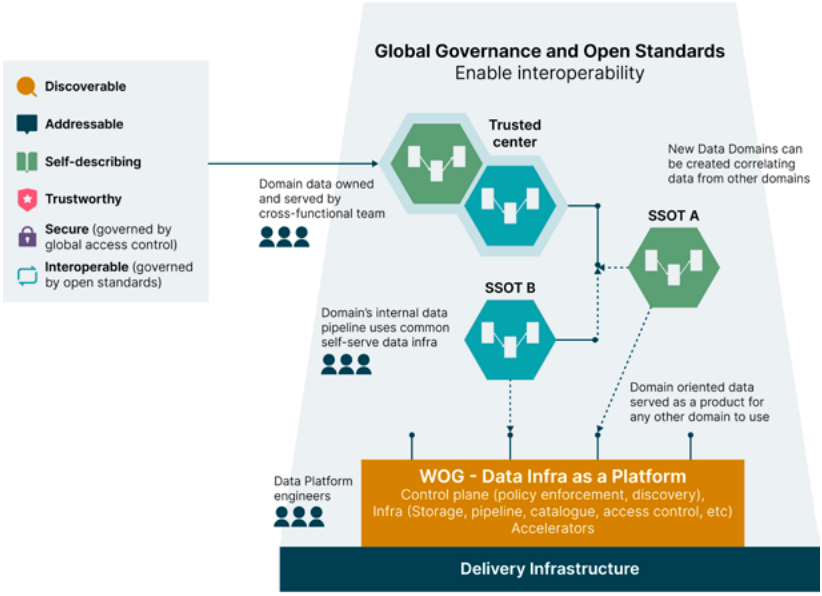

Fig 4h. Conceptual view of Data Mesh paradigm for Singapore public sector data sharing

In the next section, we take a look at an example reference architecture from AWS, on how such an approach could be implemented.

Example Reference Architecture from AWS

This section has been contributed by AWS. Reproduced with permission from AWS.© 2021, Amazon Web Services, Inc. or its affiliates. All rights reserved.

In this section, we take a look at an example reference architecture from AWS, on how such an approach could be implemented. Please note that this is an example from AWS, Thoughtworks' approach may vary, to suit individual client needs.

In the preceding section, we discussed the central concepts of the Data Mesh paradigm and showed how Data Mesh as an approach aligns with the context and objectives of data sharing in the Singapore Public Sector.

In this section, we describe a potential reference architecture for building a Data Mesh-based data and analytics infrastructure on Amazon Web Services (AWS). To realize the objectives discussed in the previous sections, we begin by defining the high-level design that builds on top of the Data Mesh pattern and the Lake House Architecture[10].

The solution also aims to integrate with Singapore Government's existing initiatives such as Whole-of-Government (WOG)'s Vault. Gov.SG.

As illustrated in the diagram below, the roles of data producers, data consumers and the Trusted Centre governance are separated. This is to highlight the key objectives of a self-service, federated governance data sharing platform. However, domain data products in their given SSOTs can assume the role of both Data Producer and Data Consumer to consume data sources, align them to specific requirements and to create, publish and operate the data products consumable through the central data catalog. Besides assuming the role of central governance for the SSOTs, the Trusted Centre also aggregates data across the SSOTs and makes these datasets available for consumption as aggregated domain data products.

[10] AWS (2021), What is a Lake House approach?

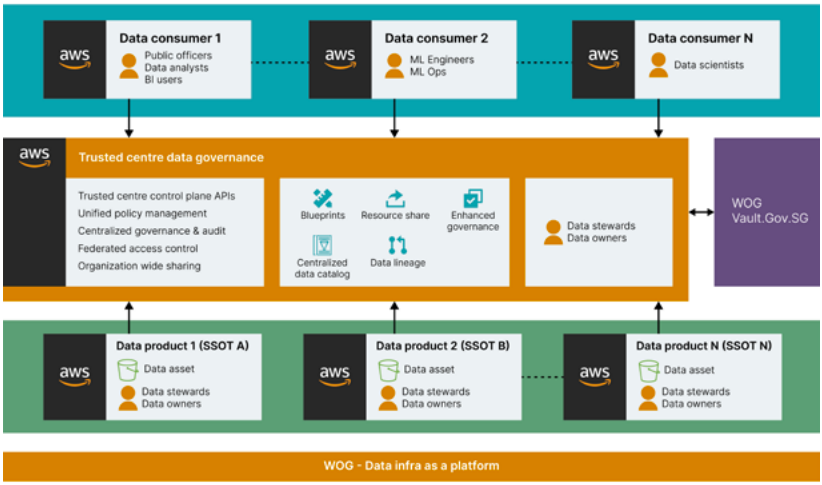

Figure 5a. High level overview of a Data Mesh design on AWS

This design illustrates a potential approach to achieve the following objectives as laid out in the previous section:

- Enabling domain driven decentralization
- Building rich domain data products
- Self-service data platform
- Federated governance

The AWS approach to building the Data Mesh identifies a set of design principles and services as building blocks of a scalable, self-service data sharing platform. The following diagram provides additional details on how AWS native services can support this design across the data producers, consumers and central governance at the Trusted Centre level.

Figure 5b. Example Data Mesh Reference Architecture from AWS[11]

Each data domain, whether a producer,consumer, or both, is responsible for its own technology stack. However, using AWS native services with the Lake House Architecture makes it possible to have a repeatable blueprint that the Trusted Centre can standardize and use to automatically provision the appropriate base stack for new domain data products as it scales the Data Mesh design. While SSOTs can build and customize on top of the base stack, having a consistent technical foundation ensures that services are well integrated, core features are supported, security and performance are baked in, and costs remain low.

AWS native services such as Amazon EMR, AWS Glue, Amazon Kinesis, Amazon Simple Storage Service (Amazon S3) and AWS Lake Formation provide the foundation for data ingestion, processing, storage and federated data cataloging at the domain data products level. This allows the SSOTs to make valuable data discoverable and available for consumption across the WOG.

[11] AWS (2021), Design a data mesh architecture using AWS Lake Formation and AWS Glue

When considering a data product for suggesting the mode of transport such as flight, bus or train based on a selected route, the data product resides in its own AWS account and uses Amazon S3 to store raw and transformed data. The data product would consume data from all other sources aligned to transport data products. The data would then be processed and transformed to a consumable form using the data product's own technology stack, or leveraging on AWS services such as AWS Glue for Extract-Transform-Load (ETL) stack, Amazon SageMaker for training and deploying AI/ML models, Amazon Macie for auto-discovery of personally identifiable information (PII). The consumable data product stored in an Amazon S3 bucket would then be made available to the federated data catalog using AWS Lake Formation.

Data consumers are granted access to the domain data products based on defined policies through the central governance at the Trusted Centre level. Public officers, subject matter experts such as data analysts, data scientists and applications can consume the data easily, making use of AWS purpose-built BI visualization, analytics and machine learning services like Amazon QuickSight, Amazon Athena, Amazon Redshift and Amazon SageMaker.

The Trusted Centre, supported by the WOG - Data Infra as a Platform, implements centralized governance, federated access controls and auditing by leveraging on the capabilities provided by AWS Lake Formation to centrally define security, governance and auditing policies, and to enforce those policies across data consumers. AWS Lake Formation provides uniform data access by creating resource shares[12] that allow the data consumers to consume the requested data directly from the domain data products.

[12] AWS (2021), AWS Lake Formation – Resource Shares

The Trusted Centre also provides the Control Plane APIs that facilitate the integration with the domain data products to effectively manage the lifecycle of the federated data catalog, ensuring that updates to source data are reflected in the data catalog as needed. The APIs also serve as the central endpoint for integrating the Trusted Centre and its SSOT domains into the WOG's Vault.Gov.SG. Under the hood, the Control Plane APIs make use of these AWS Services:

- Amazon API Gateway and AWS Lambda - providing secure endpoint for RESTful API services
- Amazon OpenSearch - providing advanced search capabilities including full-text search, suggestions, ranking evaluation and geospatial search
- Amazon DynamoDB - storing data contracts, sample datasets, data glossary and audit information
- Amazon Neptune - storing comprehensive data lineage in a high performance graph database

For accelerated adoption, consistency and standardization, the Control Plane APIs can be built based on a common framework while allowing scaling and customization for domain-specific requirements.

The common framework can be developed by a central team at the WOG level to embody the overarching central control policies for strategic alignment and smooth iterations.

As AWS innovates and adds new capabilities and services, the approach also allows for the flexibility to leverage the latest AWS innovation available to further reduce custom engineering effort.

Implementing a data mesh on AWS is made simple by using AWS managed services. By following the best practices of AWS Well Architected Framework[13], a well understood, performant, scalable, and cost-effective solution can be built to govern, integrate, prepare, and serve data.

[13] AWS (2021), Well-Architected Framework

Summary

Effective, trustable data sharing is possible, not only with technological changes but a synchronized evolution of people, processes and technology is necessary for implementing the Data Mesh. It provides a structured way of integrating people and processes with technology and evolves over time based on a feedback driven approach.

This also pushes for a shift-left mentality, placing accountability upstream on data producers, the agencies, to own and publish good quality, harmonized data. Consequently, more autonomy is achieved at the agency level, empowering them to accelerate and push the boundaries of data sharing. This paradigm can be extended to enable sharing within an agency as well, between the various departments.

The Lake House Architecture supported by the capabilities of AWS Lake Formation and other AWS managed services provides an ideal foundation towards implementing a Data Mesh-based data sharing and analytics infrastructure, providing a proven blueprint and design pattern to accelerate the delivery of domain data products across the Singapore Public Sector.

By realizing the objectives set out in this whitepaper, the Singapore Public Sector can achieve a quantum leap in the speed and scale at which data can be harnessed for insights and be integrated into business processes and services offered to the public and industries, ultimately meeting the goal of making core government data assets discoverable and accessible in a timely fashion.

References

1. Daniel Lim Yew Ma (8 Aug 2019), Bringing Data into the Heart of Digital Government, https://www.csc.gov.sg/articles/bring-data-in-the-heart-of-digital-government
2. IMDA (15 Oct 2020), Trusted Data Sharing Framework, https://www.imda.gov.sg/-/media/Imda/Files/Programme/AI-Data-Innovation/Trusted-Data-Sharing-Framework.pdf
3. GOVTECH SINGAPORE (Dec 2020), Digital Government Blueprint, https://www.tech.gov.sg/files/media/corporate-publications/dgb-public-document_30dec20.pdf
4. Thoughtworks, Whitepaper: The Data Mesh Shift, https://www.thoughtworks.com/ebook/data-mesh
5. Zhamak Dehgani (20 May 2019), How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh, https://martinfowler.com/articles/data-monolith-to-mesh.html
6. NewVantagePartners (2021), The Journey to Becoming Data-Driven: A Progress Report on the State of Corporate Data Initiatives, https://c6abb8db-514c-4f5b-b5a1-fc710f1e464e.filesusr.com/ugd/e5361a_76709448ddc6490981f0cbea42d51508.pdf
7. Singapore Government Developer Portal, Instruction Manual (IM8), https://www.developer.tech.gov.sg/guidelines/standards-and-best-practices/im8
8. Personal Data Protection Commission, Personal Data Protection Act, https://www.pdpc.gov.sg/Overview-of-PDPA/The-Legislation/Personal-Data-Protection-Act
9. Thoughtworks (May 2018), Technology Radar, Architectural Fitness Function, https://www.thoughtworks.com/radar/techniques/architectural-fitness-function
10. AWS (2021), What is a Lake House approach?, https://aws.amazon.com/big-data/datalakes-and-analytics/data-lake-house

11. AWS (2021), Design a data mesh architecture using AWS Lake Formation and AWS Glue, https://aws.amazon.com/blogs/big-data/design-a-data-mesh-architecture-using-aws-lake-formation-and-aws-glue
12. AWS (2021), AWS Lake Formation – Resource Shares, https://docs.aws.amazon.com/lake-formation/latest/dg/regranting-shared-resources.html
13. AWS (2021), Well-Architected Framework, https://aws.amazon.com/architecture/well-architected

AWS Product references

- Amazon EMR, https://aws.amazon.com/emr/
- AWS Glue, https://aws.amazon.com/glue/
- Amazon Kinesis, https://aws.amazon.com/kinesis/
- Amazon Simple Storage Service, https://aws.amazon.com/s3/
- AWS Lake Formation, https://aws.amazon.com/lake-formation/
- Amazon QuickSight, https://aws.amazon.com/quicksight/
- Amazon Athena, https://aws.amazon.com/athena/
- Amazon Redshift, https://aws.amazon.com/redshift/
- Amazon SageMaker, https://aws.amazon.com/sagemaker/
- Amazon API Gateway, https://aws.amazon.com/api-gateway/
  AWS Lambda, https://aws.amazon.com/lambda/
- Amazon OpenSearch, https://aws.amazon.com/opensearch-service/
- Amazon DynamoDB, https://aws.amazon.com/dynamodb/
- Amazon Neptune, https://aws.amazon.com/neptune/

About the authors

Sowmya Ganapathi Krishnan

Lead Consultant, Data Engineer, Thoughtworks
Sowmya has more than 12 years of experience in the industry, spanning building core infrastructure frameworks, low-latency financial exchange connectivity systems, writing data ingestion pipelines to architecting and building modern data platforms.

For the last 7 years, she has been focused in the data architecture space, solving data engineering problems using the big data stack. She recently led a data platform project in the Singapore public sector. Sowmya is very passionate about solving social problems leveraging the power of data, wherever possible.

Atif Akhtar

Lead Consultant, Data Engineer, Thoughtworks
Atif plays the role of a data consultant in ThoughtWorks with more than 10 years of experience in various capacities building large-scale distributed data processing systems, delivery infrastructure automation and federated blockchain platforms.

He has been part of the leadership team solutioning and implementing one of the largest data mesh implementation projects and has worked on and led various other holistic data platform implementations driving high impact business relevant solutions.

Notices

This document is provided for informational purposes only. It represents current product offerings and practices as of the date of issue of this document, which are subject to change without notice. You are responsible for making your own independent assessment of the information in this document and any use of products or services mentioned, each of which is provided “as is” without warranty of any kind, whether expressed or implied. This document does not create any warranties, representations, contractual commitments, conditions or assurances from Thoughtworks, its affiliates, suppliers or licensors.

Ready to start your innovation journey?

Thoughtworks and AWS are here to help you along your innovation journey. Contact Thoughtworks at partnerships@thoughtworks.com to discuss your readiness to take the first step.

/thoughtworks

aws partner network
Advanced
Consulting
Partner