LTIMindtree

POV

Data Mesh and its Applications in Data & Analytics

Authored by:
Christeena Uzhuthuval

Introduction

"Earlier, our production workload used to run for about 2 hours a day. But several bottlenecks have caused it to increase to over 6 hours daily in recent weeks. We have set up a war room with the ETL, BI, DBA, Infra, and Source teams to identify the root cause and assign accountability to the concerned team. But nobody seems to own up to the task, because it does not directly impact them."

Do these statements ring a bell? Are they something you hear in your regular projects which follow the traditional waterfall model for implementation?

The teams involved in the above scenario are divided based on technologies or technology solutions, which means they do not necessarily understand the complexities of each other's work. This has been an issue in many organizations as of late.

Data and analytics have long been an enterprise-driven centralized activity revolving around data warehouses, data lakes, and, of late, lakehouse architectures. Data mesh came in as a sociotechnical paradigm: a shift in thinking which arose due to complexities in the existing enterprise solutions that could not satisfy the demands of growing organizations.

So, what exactly is data mesh and how does it fit into our current enterprise models and help us solve our problems? To answer this, let us delve deeper into its beginnings.

Origin of Data Mesh

Data mesh as a concept came into prominence after Zhamak Dehghani coined the term in 2019 while being engaged as a principal consultant at a technology company. As Zhamak explains in her pioneering book 'Data Mesh - Delivering Data-Driven Value at Scale', "data mesh is a **decentralized sociotechnical approach** to share, access, and manage analytical data in complex and large-scale environments." Does that sound like a whole lot of jargon? If yes, no problem, we will go through the concepts in the subsequent sections.

Though data mesh is a novel approach to thinking, the underlying principles are practices that have evolved over the last two decades. This article attributes much of its insights to the ideas expressed in the above-mentioned book.

Current Day Data Architecture and Solutions

Traditional or modern data architectures typically have data brought in from various sources into a centralized location. Different ETL or ingestion tools connect to the various data repositories, which could vary from on-premises databases to streaming IoT inputs and social media outlets. All this data is brought into a data lake or warehouse and cleansed, curated, and transformed to satisfy the current business requirements. A semantic layer is exposed, which would satisfy the end-user reporting and dashboarding needs. In such a scenario, teams are mostly technology-focused, limited to one area of expertise.

The below gure (Fig 1.) depicts a modern lakehouse architecture in Azure cloud. This wireframe would be somewhat similar to any other cloud, except for the technology-specific tools used, which would change across vendors.

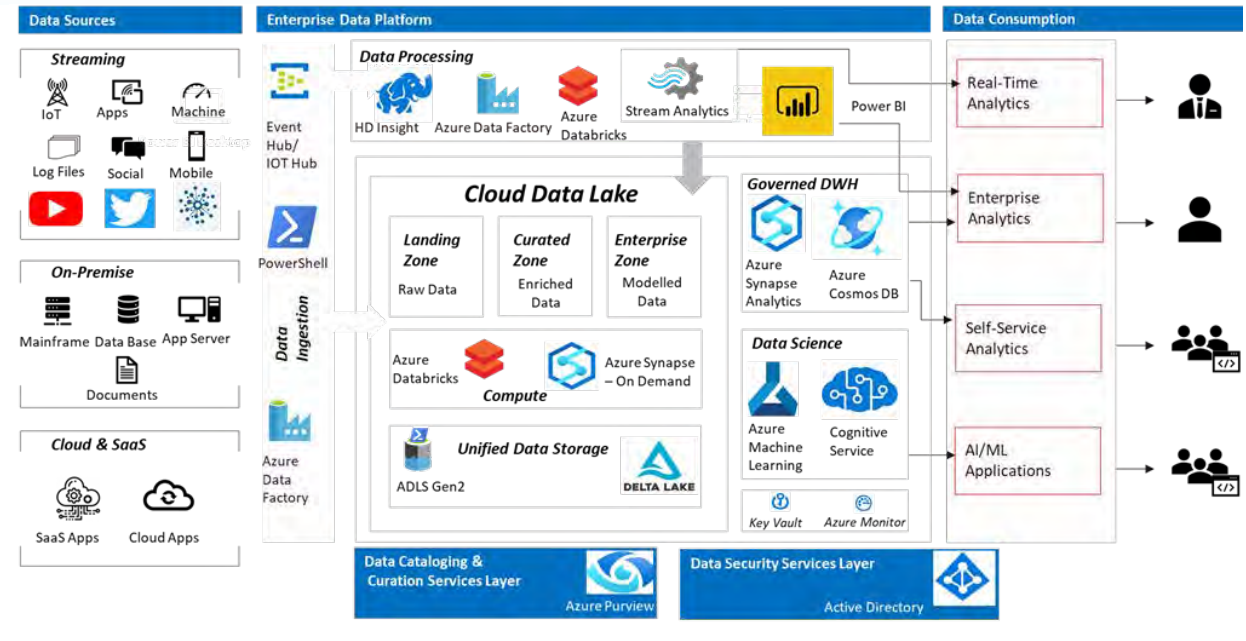

Figure 1: Sample data lakehouse architecture with azure

This architecture works fine when the volume of data is in the gigabytes. Data is all in a centralized place, easily accessible to the development teams, data scientists, and analysts to use in a very easy and efficient way

Need for Change

As organizations grow, business requirements and expectations scale very rapidly. Growing data volumes introduce complexities in implementing solutions since executing a change to a flow involves multiple data units and teams. Development teams feel hesitant to pick up new changes owing to the complexity of testing the entire process. There could arise a situation where individual process owners try to push the responsibility of the change to downstream or consuming parties, citing the current process or work overload.

New Way of Thinking

Data mesh is a distributed data architecture, and the ownership model is aligned to business units (also referred to as domains) while having data shared as a viable product. Data is no longer a by-product of your transformations; instead, you make decisions based on your data. It's a complete paradigm shift from the approach mentioned above. Let's go through some key principles of data mesh in detail.

- **Decentralized ownership of data** ensures that people and teams nearest to the data, i.e., the first-class users are accountable for providing quality data. Identifying these boundaries could be challenging for traditional architec-ture teams. The data mesh approach suggests dividing the responsibility based on the business domains. This makes each data team responsible for sharing data in an enriched and useful way. The business domain should have a specific business objective and an outcome it strives to achieve. This method can also be called as 'domain-driven design' (DDD).
- **Data is considered a product** instead of treating it as a by-product of running transformations and code pipelines or merely as an asset. This is done by applying product thinking to domain-oriented data to make it valuable and feasible to all its consumer personas. This brings in the concept of data products: a published data set that is valuable and directly usable by the downstream to generate business value. Data and the code which maintains it are considered one autonomous unit, and decisions are data-driven.
- **Self-serving data platforms** focus on making the entire data-sharing process across teams smooth and efficient. The platform integrate both operational and analytical functionalities and serves autonomous domain teams. It can scale out in a decentralized way. It favors decentralized technologies, and the end data products are interoperable. Users of this data should be able to discover and use the data products seamlessly without dependency on

centralized teams. Cross-functional teams need to share data in an easy and efficient way.

- **Federated governance of data with computational policies** in place assures that all independent data products are secure, trusted, and deliver value through correlation. With data mesh governance, changes to the ecosystem are embraced, and security, compliance, quality, and usability of data are computed in an automated way. Computational policies are embedded early in the life cycle of each data product and monitored throughout. Though each domain is autonomous, it must comply with global organizational compliance and interoperability standards.
- **Interoperability through a distributed mesh** helps in connecting data products that can be accessed using standard APIs rather than collecting data in monolithic warehouses and data lakes.

Fig 2. below shows the basic attributes given by Zhamak that every data product must have to be considered useful.

Figure 2: Usability attributes of data products as given by zhamak (Dautnivs)

Every domain is responsible for its analytical data ownership and lifecycle management. This is made possible as accountability of the data is on the producers or owners of data. Every domain will be part of a distributed architecture, maintaining its own datasets, code, and data policies and having specific business goals. Individuals who best know and understand the nitty-gritty of data in their domain are responsible for managing it and working together with other domains using well-defined APIs. To treat their data as a first-class p oduct, organizations need to facilitate self-serve platforms for building and publishing data products backed by a steady federated governance structure. Each data product is a self-governing unit and is managed independently by the data owner.

Modern Architecture with Data Mesh

Now that we have seen and understood the data mesh approach, let's consider the changes we can bring about in our current architectures to make these agile and respond gracefully to changing demands in the ever-growing business landscape. This will help increase the data-to-investment ratio, and at the same time, sustain agility in the face of growth.

Let us consider an insurance sector use case. An insurance organization has many business domains, such as policies, claims, and premiums. The data is brought in by multiple sources and ingested into data marts, data lakes, and data warehouses. Following the data mesh approach, we can break down the implementation to follow an agile way of work and release.

Each domain team will have an ETL developer, tester, analyst, data scientist, data modeler, DevOps engineer, data product owner, and so on. Fig 3. below depicts a sample data mesh architecture for the policy domain. Other domains would follow a similar structure. The different processes are explained below. Snowflake has been conside ed as the target data cloud here. Any cloud vendor can be used and would have a similar structure and alignment in terms of lifecycle processes.

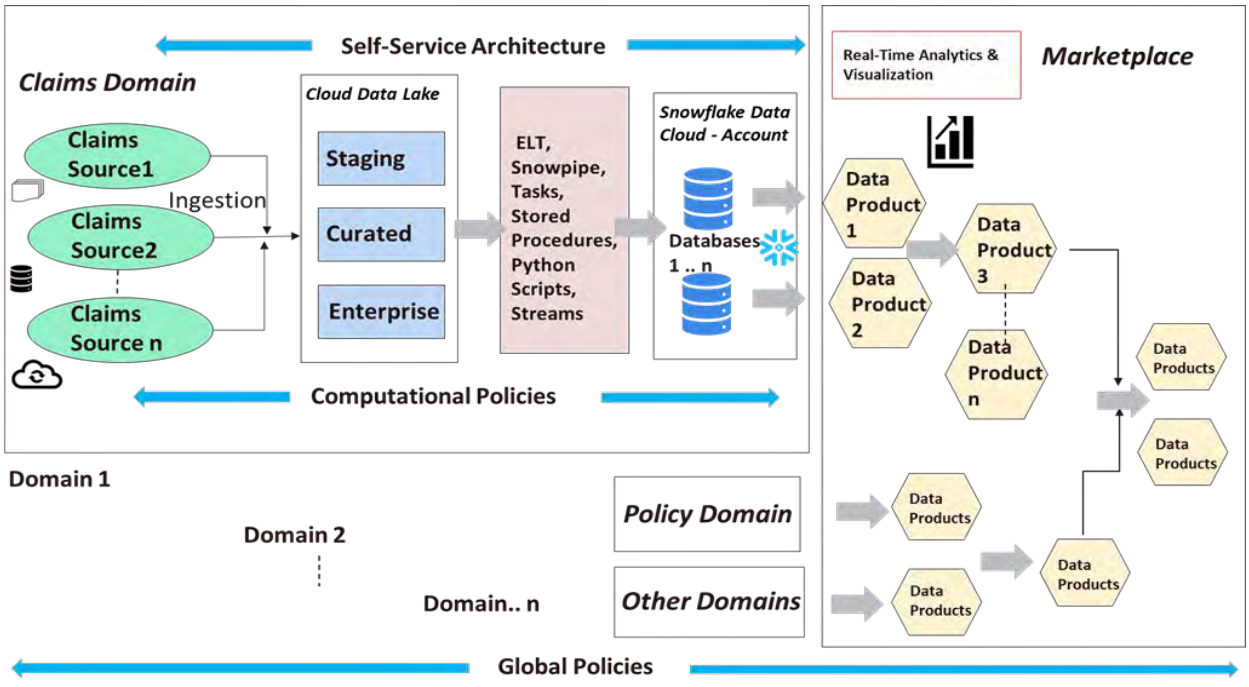

Figure 3: Implementing data mesh on snowflake data cloud

- In the above figure, I have considered Snowflake as the data cloud, with **each account corresponding to a single domain.** Multiple databases within the Snowflake account can correspond to multiple data products offered by the domain, which are easily accessible by other domains. All the Snowflake accounts corresponding to the different domains fall under one Snowflake organization.
- Each domain has a dedicated data product owner who focuses on the data shared by that domain. The business data owners will manage the entire data lifecycle, thus providing a single source of truth for that domain. Each team has long-term accountability for providing data that is easily discoverable, understandable, accessible, and usable, and is known as data products. This will be facilitated by the self-serve infrastructure team and governed using computational policies created and provisioned by the policy steward and the domain provisioner, respectively.
- Data products can be shared and accessed automatically with other data teams by registering them on the marketplace which will be the global data discovery tool. These products assure a set of standards and need to have clear documentation. The marketplace will have provisions for registering your data product. The following details will help increase the score for the data product – an overview, discoverability, trust, usability, observability, security, cost, etc.

- The more a data product is consumed and used, the better usability score it will have. This helps in tracking data products and removing or reevaluating the data products which are not being used or consumed by others.
- The intermediary data products could be the database or database objects, or reports run on any of the data products.
- Changes to data products can be pushed periodically using the DevOps process within the team. Access would be controlled through federated governance in an automated fashion.

Conclusion

Data mesh is a new and innovative approach that aligns data and business closer than ever to harness values and insights from analytical data at scale. Analytics data could have varied use cases. Changing from traditional architecture to a data mesh approach at an organizational level could be difficult at the beginning. B eak it down into small steps that can be taken, keeping in mind the bigger picture. **Start with one small business domain** and establish the new flow independently without interrupting the existing flow. Once it is up and running, gradually move on to the next business function. The new architecture also needs to be able to support the continuous delivery of increasing new features into existing data products that evolve from the ever-changing business.

With data mesh, any data team can quickly get access to any data, across the organization. Data teams need to act quickly to growing demands with agility to facilitate near real-time decision-making. Change and volatility need to be embraced and not considered taxing. Autonomous individual teams following the data mesh approach can help organizations to fully become data-driven, making the best use of data and analytics. No data is irrelevant when used to its full potential!

References

1. ‘Data Mesh - Delivering Data-Driven Value at Scale’ book by Zhamak Dehghani
2. https://www.datamesh-architecture.com/
3. Discussions with LTI Data Mesh Team

About the Author

Christeena Uzhuthuval is presently Technical Lead for PolarSled Team and part of Snowflake COE at TI in a Specialist role. She has more than 10 years of experience in Data & Analytics, is passionate about Data Cloud, and likes to keep herself updated on emerging technologies in the Data domain. She likes spending her free time reading novels, engaging in some sports, or listening to music.