spectra

by Fosfor

Point of View

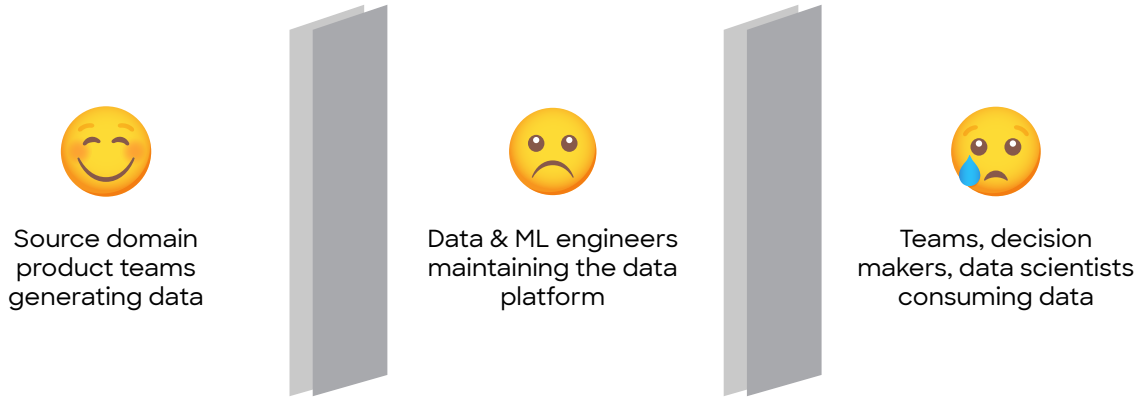
Data Mesh and Its Impact on Data Platform Providers

Introduction

We have been dealing with data for decades, and there has been significant innovation in the underlying technologies over the years. But of late, there has suddenly been a huge buzz around "Data Mesh." To date, legacy data warehouses have served as a single-source-of-truth for all business analytics needs, but now the advent of data lakes has solved what was wrong with data warehouses. Data lakes not only fulfil the requirement of today's diverse data needs in terms of variety (structured, unstructured, semi-structured), velocity (real-time vs. batch), volume (PB vs. GB) but also in terms of veracity (data integrity when using data from multiple sources, timeframes, and accessible to a multitude of stakeholders, users, and consumers).

While data warehouses are good from a historical and analytical perspective and typically explain past trends, they are seriously constrained while providing a future perspective – especially when it comes to predictive and prescriptive analytics, a key requirement for data-driven decisions (DDD). What data lakes do is facilitate and make the life of data scientists a bit easier by retaining the integrity of original transactional data from online transaction processing (OLTP) systems/sources. This data processing/ massaging/ transformation has shifted to last-mile activity (Extract-Load-Transform (ELT) vs. erstwhile Extract-Transform-Load (ETL)) in the data pipeline.

One important fact that needs to be addressed is the core challenge that creates the need to look for alternative solutions like Data Mesh, as depicted in the diagram below.



Ref: Zhamak's blog and Spark+AI summit 2020

It boils down to whether data ownership should remain with decentralized domain teams who have the required skillset and knowledge regarding the value and context associated with data OR if it should be delegated to a technically competent centralized layer of data engineers. Both approaches have their pros and cons. Decentralized ownership with domain teams who are the original producers and aware of the data sources makes more sense because eventually, quality of outcome (insight/ prediction/ explanation/ decision) is underpinned heavily on the inherent meaning of data. They are the ones who understand the nuances, values, and impact of this data. If this data is handed off to expert data engineers in its raw form, they can manipulate the data in every possible way to serve the needs of the consuming teams (data scientists, analysts, etc.). When this happens somewhere along the way, the value of the data diminishes with this engineering transformation lacking depth of domain understanding. In turn, this impacts the result that "Models" generate in terms of accuracy and meeting business objectives.

To be fair, there is merit in the prevailing centralised architecture that we have today. Knowledge building and cross leveraging the scarce talent pool of data engineers and SMEs is far more effective in this setup. Economies of scale are a default benefit of such centralised ecosystems. Also, central governance is much easier to implement. It can also be argued that running the operation and core transactional business requires a different line of thinking of what is required for business analytics. Hence, it is better to have separate teams own these responsibilities.

Having represented both arguments, the one fact that cannot be denied is that there are multiple limitations in the current architectural approach, whether it is a data warehouse or a data lake. Both are centralised and can thus cause a bottleneck for the central data team. Scalability and autonomy are hampered in this architecture, and this is the exact reason for the evolution of Microservices-based architecture in the application space. It has been proved beyond doubt that legacy monolithic architecture is not the solution for the diverse, distributed, and hyper-scalable requirements of today's agile business. The same set of core principles are applicable for the data analytics space as well. Monolithic is synonymous with sluggishness in transaction systems. Hence, the question is whether it is different for analytics systems. Therefore, here we see the need for the proposed paradigm shift that the Data Mesh approach brings.

Principles of Data Mesh

There are four founding principles of Data Mesh as described by lead thinker Zhamak D. (Ref: https://martinfowler.com/articles/data-mesh-principles.html). While it is still early and most data platform providers are grappling with their own interpretation of mapping these founding principles with platform features, a need arises to demystify what they potentially mean for the data infrastructure platform.

Domain Oriented Decentralization

- Multitenancy where common platforms can be leveraged across different domains / groups within the enterprise without compromising on isolation and separation
- Data security and other enabling tools to achieve various compliances like GDPR, CCPA, HIPPA, etc.
- Access control and governance (RBAC, IAM)
- Polyglot capability and flexibility to use popular languages/ engines/ runtimes
- End to end DataOps and MLOps capability to facilitate self-sufficiency among domain teams
- Citizen developer LC / NC enablement and tools

Data as a Product

- Ability to publish data for consumption (as API) in a secure, managed, and governed manner
- Tools to clearly define and publish metadata that is discoverable (Data Catalog)
- Tools to enforce data quality and metrics around that
- Traceability of data through data lineage
- Measuring the success of the Data Product like any other product in terms of CSAT
- Governance and control

Self-Serve Data Infra as a Platform

- Domain agnostic Platform as a Service (PaaS)
- Easily provision underlying infrastructure using this platform, ensuring domain teams can focus on driving value from data
- Scalable, Secure, Hybrid, Resilient, and any other Non-Functional Requirement (NFR) that you want to add
- A Fast-learning curve that helps domain teams in quick ramp-up and onboarding team members

Federated Governance

- Codified standards and best practices to a possible extent, even during the evolutionary process
- Easy and pluggable capability for authentication and authorization
- Federated identity and access management
- Tools for Key Performance Indicators (KPI) and Service Level Objectives (SLO)-based metrics for homogeneous implementation across domains
- Alerts and notifications

Conclusion

Any fundamental shift of this nature that has ramifications across the industry will be evolutionary and slow. Huge investment in existing environments and the current, established ways of working will be a major challenge to overcome. It will be interesting to see how the different players in this ecosystem respond to the impending changes that are truly transformational.

By Ritu Raj Tripathi

Director
Platform Engineering

Ritu Raj Tripathi is the Director, leading the engineering platform at Fosfor Spectra. He comes with over two decades of extensive experience in Information Technology services and products. His expertise includes data and analytics, digital transformation, integration, API, Microservices, cloud-native technologies, App Modernization, distributed architecture, and program management. He holds a BE from NIT Allahabad (MNNIT) and executive management (EPLM) from IIM Calcutta.
Ritu Raj Tripathi is the Director, leading the engineering platform at Fosfor Spectra. He comes with over two decades of extensive experience in Information Technology services and products. His expertise includes data and analytics, digital transformation, integration, API, Microservices, cloud-native technologies, App Modernization, distributed architecture, and program management. He holds a BE from NIT Allahabad (MNNIT) and executive management (EPLM) from IIM Calcutta.

The Fosfor Product Suite is the only end-to-end suite for optimizing all aspects of the data-to-decisions lifecycle. Fosfor helps you make better decisions, ensuring you have the right data in more hands in the fastest time possible. The Fosfor Product Suite is made up of Spectra, a comprehensive DataOps platform; Optic, a data fabric to facilitate data discovery-to-consumption journeys; Refract, a Data Science and MLOps platform; Aspect, a no-code unstructured data processing platform; and Lumin, an augmented analytics platform. Taken together, the Fosfor suite helps businesses discover the hidden value in their data. The Fosfor Data Products Unit is part of LTI, a global technology consulting and digital solutions company with hundreds of clients and operations in 31 countries. For more information, visit Fosfor.com.

FOSFOR by LTI