DATA MESH TURNS CENTRALIZED OWNED DATA LAKES INTO THE VILLAIN

HOW CAN A DATA MESH HELP ENTERPRISES DELIVER DOMAIN-CENTRIC BI CAPABILITIES?

This article is part of Capgemini's Applications Unleashed 2022 Report which can be downloaded here.

HIGHLIGHTS

....

- The data industry went through a serious transformation.

....

- However, the mental model didn't change and with that kept inherently failure modes in place.

....

- The data lake is no longer the centrepiece of the overall architecture.

....

- Dark data is never leaving the operational ecosystem and is not used for other purposes.

....

- Data mesh puts the business in control of how data is being used within a domain.

....

To become a data-driven intelligently empowered company, organizations need to revise the current mental model of data.

Considering ever-exceeding volumes of data being collected, the priority of companies around the globe has been to unleash the power of data and harvest the benefits of being an intelligent company. It is no surprise, in the light of data being coined to be the new oil (2006), that companies create centralized data platforms to counterbalance the risk of having accidental silos of unreachable data and missing out on this promise. But companies in the quest for greatness were deluded to create failure modes, anti-patterns, and a disconnect from the business. The article explains to you how a data mesh can help your organization to invert the current mental model on data, which will bring authoritative and filtered data into our lives leading to trustworthy sources so we can make fast and accurate decisions.

WHERE DO WE COME FROM?

To understand how we got to the current tight spot we are in, we need to take a walk down memory lane, looking at the predecessors of the current generation of data architecture. In the past first-generation data platforms were humongous proprietary data warehouses (DW) to meet data regulation and business needs, with a large capital expense (CAPEX) beforehand and bespoke ETL (Extract-Transform-Load) jobs to get it all running. The scale crisis resulting in more and more data coming in has led organizations to move away from overflowing data warehouses to second-generation big data ecosystems with a data lake as the golden child. A centralized data (lake) platform with agnostic data ownership disconnects

them from business and hyper-specialized teams owning, handling, and creating the data lake "monster" in its current form. With the introduction of hyperscalers and moving to the cloud a third generation of data-lakes modernize the data architecture introducing real-time analytics (streaming) while simultaneously moving away from beforehand costs to operational expenses. Even though the data industry went through quite some transformation the mental model didn't change and with that kept inherently failure modes in place.

THE FLAWS OF THE CURRENT DATA LAKE ARCHITECTURE

The current mental model has two pressure points that often lead to failure. Firstly, to prevent the creation of accidental silos of unreachable data and the fear of missing out on the data being the new oil, organizations answered by centralization and harmonized their data under one centralized control and its ownership with one hyper-specialized team. This agnostic and centralized take on data ownership is a pressure point, creating data engineers that have been siloed and disconnected from the business, that don't allow the exchange of knowledge around what the data needs to be. Secondly, once you have a centralized piece of architecture, your architecture and your organizational structure are going to mirror each other which is called Conway's law [1]. So, in that case, you end up with a centralized kind of platform and a centralized team responsible for it. Resulting in multiple bottlenecks like highly coupled pipelines, long response times, and reduced ability to consume data and adding new sources. This should be considered an anti-pattern because well-designed architecture should accelerate change over time.

---

1 Conway's law, https://en.wikipedia.org/wiki/Conway%27s_law
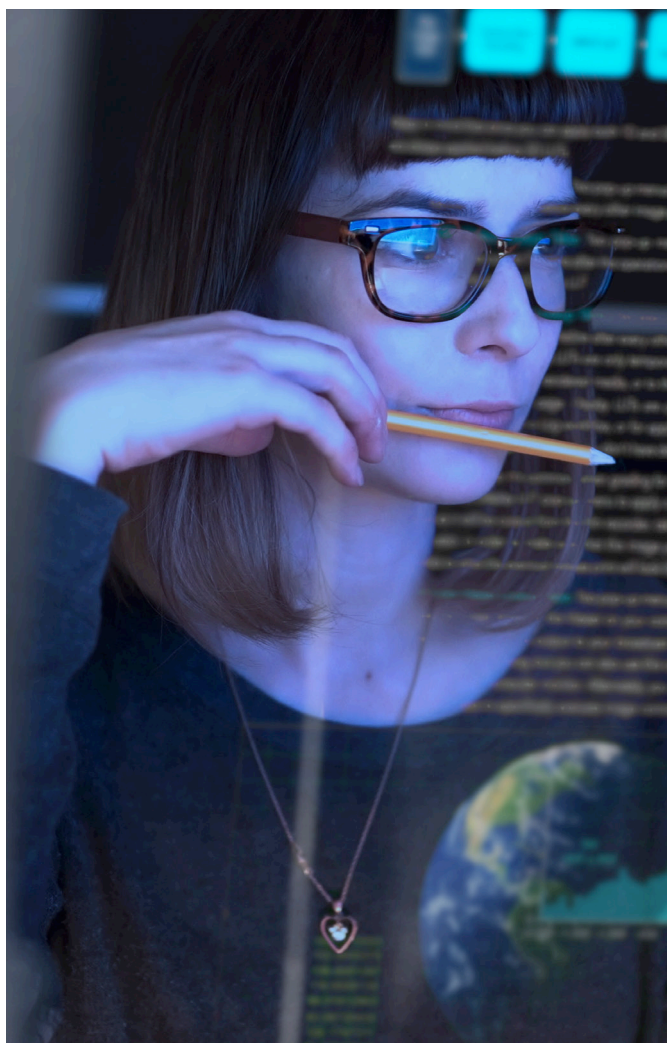
Recently, with trends like hyper automation and business increasingly being digitized, analytics and AI can be used to optimize and automate operational processes

THE TWO REALMS OF DATA

We need to shift from an it-centric view, with as a centrepiece the data lake, to a business-orientated focus. Changing the current mental model is where the data mesh comes in. But before that, we require some basic understanding of the different types of data.

Today's landscape is divided into two levels of existence, operational- and analytics data. Operational data is the more mature one, largely driven by a modern microservices architecture. Serving the needs of applications running the business. Analytics data provide an aggregated view of facts of business over time, modelled to get business intelligence, and future insights and feed analytic reports to management to steer the organization on its path of success. Management access to analytics data remains a point of friction. This becomes apparent when looking at the process and transactional data that is collected, processed, and stored during regular business activities, of which 93% never leaves the operational eco-system and generally fails to be converted to analytics data and be used for date-driven decision-making according to IBM [2]. This is called 'Dark data[3]' by Gartner. Moving data from operational applications to analytics solutions was usually a one-way street. Recently, with trends like hyper automation and business increasingly being digitized, analytics and AI can be used to optimize and automate operational processes. The two worlds need to converge into one, for which the data mesh is essential.

NEW ENTERPRISE DATA ARCHITECTURE

Applying learnings from the past decade in building distributed architecture at scale, applied to data is what is called data mesh. The primary focus of data mesh is on analytics data. One of the concepts from the realm of distributed architecture that is vital for a data mesh is domain-driven design. This means that architectural structures are aligned with the domain in which they are used[4].

THE SHIFT FROM IT-CENTRIC TO BUSINESS SERVICE ORIENTATED

These insights cause a paradigm shift from a centralized data lake to an eco-system of data products with a business-orientated focus and provide an operational shift as well. A well-defined data product should be recognizable by businesses in the context of their domain and domain language, changing from an inside-out view of datasets, being interpreted by siloed big data engineers to an outside-in data-product view. This change from a product to a service focus is also the underlying concept of innovating banking. Customers are more focused on the service it delivers than the underlying product that the financial institution provides.

2   IBM, https://www.ibm.com/blogs/cloud-archive/2015/11/future-of-cognitive-computing/
3   Dark data, https://www.gartner.com/en/information-technology/glossary/dark-data
4   Domain Driven Design is Key to adopting digital transformation, Page 16, Application Unleashed 2021. https://www.capgemini.com/nl-nl/wp-content/uploads/sites/7/2021/09/Applications-Unleashed-Report.pdf

They don't want a mortgage, but they want somewhere to live, not a loan but a boat or car. An understanding by the domain data teams provides a common language and context to the data engineers to understand how the data is used within the domain. This provides guidance on uncovering and unlocking "dark data" and converting it to analytic data.

Consider as an example, a digital customer experience platform for a financial Institute with different domain-oriented data-products. One of the critical domains is 'contracts'. Which contract does the customer have, when was it requested, and by whom? This key domain has different consumers in the organization. For example, near real-time consumers that are interested in the buyer's journey or somebody who has a mortgage who could dynamically be offered home insurance, or an incoming disgruntled customer support call to provide a loyalty program. The reason this dataset can provide the best user experience for consumers is that the way of doing business is understood by the domain.

FLAWS IN THE CURRENT MENTAL MODEL RESOLVED

Moving to a data mesh changes the perspective on the two failure modes present in the current data architecture. By moving away from the data lake as the centerpiece of the architecture to being simply nodes in the mesh. Moving away from the data lake supporting data science access patterns and data warehouses supporting analytical and business intelligence reporting access patterns. It allows distributing the data products around the business domains and their ownership. An operational shift will take place to independent, cross-functional teams on the domain level. This provides flexibility, autonomy, and a shift left to the teams in the same way as in distributed architecture. To facilitate this autonomy on the team level, the smallest unit in data architecture - the data pipeline - must be decentralized and oriented in such a way that the teams have end-to-end control. A domain data product owner is responsible for ensuring the data is aligned with the strategic goals of the domain and that data is delivered as a product. The data products support the business services and are evaluated on data quality, lead time of consuming data, and data user satisfaction, which can also come from a customer of the organization through NPS (net promotor score).

THE PARADIGM SHIFT TOWARD A DATA MESH

Let's bring it all together. We looked at the characteristics of the current data platforms: centralized, monolithic with highly coupled pipelines, operated by silos of hyper-specialized data engineers. Next, we introduced the data mesh that inverts the current mental model from an IT-centric to a business-orientated approach, in which the learnings from distributed architecture are applied to analytical data. It introduces distributed data products orientated around domains owned by independent cross-functional teams and governed by a data product owner. The model also brings analytics data and operational data closer together under the same domain while it also respecting their underpinning technical differences. Reducing the amount of dark data that exists within the organization. With the data lake being a technical implementation instead of a centerpiece of the overall architecture, it can be used in the way it was intended: data lakes are used for single domains, creating a water garden of domains within your organization. This paradigm shift democratizes data to business and puts the organization on the road to becoming the data-driven intelligent, empowered company they always wanted.

ABOUT THE AUTHOR

Guido Bosch
Solution Architect

Guido is a Solution Architect with a focus on combining a data-driven approach with Hyper Automation and Customer Engagement. With over 8 years of experience in the field of financial services, he performs a wide variety of projects in security, data, cloud, and low code.

guido.bosch@capgemini.com

https://www.linkedin.com/in/guido-bosch/

About Capgemini

Capgemini is a global leader in partnering with companies to transform and manage their business by harnessing the power of technology. The Group is guided everyday by its purpose of unleashing human energy through technology for an inclusive and sustainable future. It is a responsible and diverse organization of 325,000 team members in nearly 50 countries. With its strong 55 year heritage and deep industry expertise, Capgemini is trusted by its clients to address the entire breadth of their business needs, from strategy and design to operations, fueled by the fast evolving and innovative world of cloud, data, AI, connectivity, software, digital engineering and platforms. The Group reported in 2021 global revenues of €18 billion.

Get the Future You Want | www.capgemini.nl

For more information:

Capgemini Nederland B.V. P.O. Box 2575, 3500 GN Utrecht Tel. + 31 30 203 05 00 www.capgemini.nl