FINBRIDGE

Data Mesh: Eine dezentrale Datenarchitektur

Ilja Jost

04. Juli 2023

Data Mesh: Eine dezentrale Datenarchitektur

In Finanzinstituten wie Banken und Versicherungen wird das Sammeln, Verwalten und Analysieren von Daten immer wichtiger. Die Datenmenge, die von diesen Unternehmen generiert wird, wächst exponentiell und es wird zunehmend schwieriger, sie effektiv zu nutzen, um Geschäftsentscheidungen zu treffen und wettbewerbsfähig zu bleiben. In diesem Kontext kann die Implementierung von Data Mesh eine wertvolle Lösung sein, um die Herausforderungen bei der Datenverwaltung zu bewältigen.

Was ist Data Mesh?

„Data Mesh ist ein dezentraler soziotechnischer Ansatz für die Bereitstellung, den Zugriff und die Verwaltung von analytischen Daten in komplexen und großen Umgebungen – innerhalb eines Unternehmens oder organisationsübergreifend." [1]

Soziotechnisch bedeutet, dass es sich hierbei nicht nur um einen rein technischen Architekturansatz handelt, sondern sich auch die organisatorischen Strukturen, Prozesse und die Kultur verändert. Data Mesh nimmt Einfluss auf folgende Bereiche:

Organisation: Die Verantwortung für die Daten liegt nicht mehr zentral bei den Spezialisten, die die Datenplattform betreiben. Es entsteht ein dezentrales Modell, das die Verantwortung für die Daten in die Domänen verlagert, aus denen die Daten stammen oder in denen sie verwendet werden.

Architektur: Die Daten werden nicht mehr über eine monolithische Datenplattform zur Verfügung gestellt, sondern in einem verteilten Mesh von Datenprodukten bereitgestellt.

Governance: Statt einem zentralisierten, operativen Top-down-Modell mit menschlichen Eingriffen wird ein föderales Modell, bei dem automatisierte Policies in das Mesh integriert sind, verwendet.

Daten: Das Wertesystem verlagert sich weg von Daten als Assets, die gesammelt werden, hin zu Daten als Produkte, die internen und externen Datennutzern zur Verfügung gestellt werden.

Infrastruktur: Die Infrastruktur wird nicht mehr durch zwei getrennte Welten für operative und analytische Systeme dargestellt, sondern durch eine integrierte Einheit abgebildet.

[1] [Deh23] Zhamak Dehghani „Data Mesh – Eine dezentrale Datenarchitektur entwerfen" O'Reilly 2023

Data Mesh kann also als Bestandteil einer unternehmensweiten Datenstrategie verstanden werden, die den Zielzustand sowohl der IT-Architektur als auch eines organisatorischen Betriebsmodells beschreibt.

Warum Data Mesh?

Um zu verstehen, wo Data Mesh ansetzt, hilft es, die bisherige Entwicklung von Datenarchitekturen, wie in Abbildung 1 dargestellt, zu betrachten. Die Anzahl an Datenquellen, das Volumen, die Geschwindigkeit und die Vielfalt an Daten nimmt stetig zu. Mit dieser Entwicklung steigt gleichzeitig die Komplexität der Datenplattformen. Für das zentrale Datenteam wird es immer schwieriger, die Daten zu händeln und die vielen analytischen Fragen des Managements und der Fachbereiche zu beantworten. Das zentrale Datenteam wird zum Bottleneck.

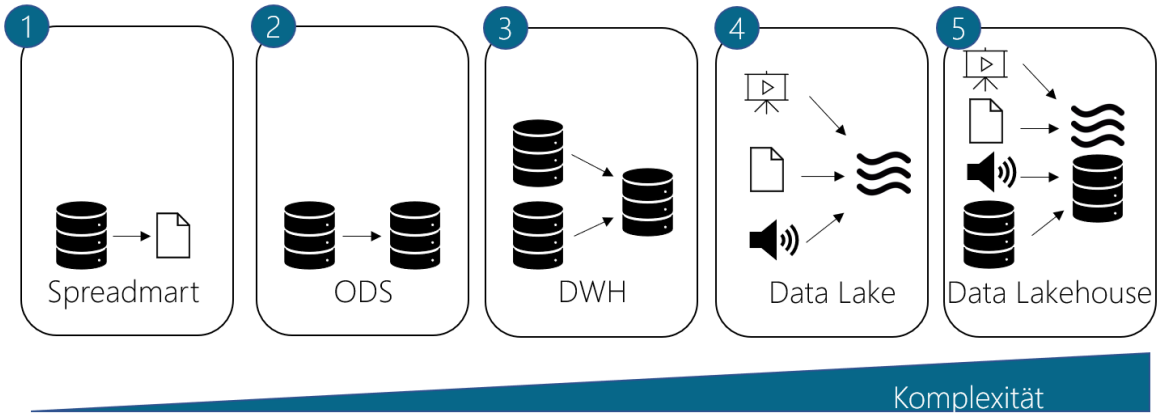

Abbildung 1: Datenarchitekturen aufsteigend nach Komplexität

(1) **Spreadmart:** Hierbei werden die Daten direkt aus dem operativen Quellsystem extrahiert und z. B. in Excel-Dateien (Spreadsheets) oder lokalen Datenbanken wie MS Access gespeichert. Die Erstellung der Spreadmarts erfolgt individuell und spezifisch, um Anforderungen eines Unternehmens oder einer Abteilung zu erfüllen. Im Laufe der Zeit werden diese Spreadmarts oft auf verschiedene Weise aktualisiert und geändert, ohne jedoch eine einheitliche und konsistente Datenstruktur zu gewährleisten. Ein Spreadmart ist i. d. R. einem Quellsystem zugeordnet.

(2) **Operational Data Store:** Der ODS speichert und integriert Daten an einem zentralen Ort. Im Gegensatz zu einem Data Warehouse (3), das Daten langfristig speichert, sind die Daten in einem ODS flüchtig. Ändert sich ein Datensatz im Quellsystem, so wird der entsprechende Datensatz auch im ODS aktualisiert. Somit sind die Daten in einem ODS zu jeder Zeit (nur) ein aktueller Snapshot der Unternehmensdaten. Die Daten werden in einer nicht-aggregierten Form abgelegt und weisen somit eine hohe Granularität auf. Ein ODS ist häufig einem oder wenigen Quellsystemen zugeordnet.

(3) **Data Warehouse:** Ein DWH dient als zentrale Datenschnittstelle, die alle Datenströme der Quellsysteme aufnehmen, integrieren, transformieren, historisieren und bereitstellen soll. Die Daten werden in einer strukturierten und häufig aggregierten Form abgelegt. Ein DWH wird oft für traditionelle BI-Anwendungen (BI = Business Intelligence) verwendet, bei denen die Datenstruktur im Voraus bekannt ist und die Daten in einem vordefinierten Schema organisiert werden.

(4) **Data Lake:** Ein Data Lake ist eine flexible, skalierbare und nicht-strukturierte bzw. semi-strukturierte Datenlagerung, die eine Vielzahl von Datenquellen aufnimmt, ohne dass im Voraus ein Schema definiert werden muss. Ein Data Lake ermöglicht es Unternehmen, große Mengen von Rohdaten aus verschiedenen Quellen aufzunehmen, ohne sich Gedanken darüber zu machen, wie sie strukturiert oder organisiert werden sollen. Data Lakes können daher sehr schnell und einfach Daten aufnehmen. Die Herausforderung besteht darin, die Daten später zu organisieren und zu strukturieren, um sie für Analysen nutzbar zu machen. Ein Data Lake wird häufig für ML-Anwendungen (ML = Machine Learning) genutzt.

(5) **Data Lakehouse**: Bei einem Data Lakehouse werden Elemente aus dem Data Warehouse und dem Data Lake kombiniert. Hier werden aus einer Vielzahl von Systemen strukturierte, semi-strukturierte und unstrukturierte Daten abgelegt. Die Daten können für klassische BI-Anwendungen und für moderne ML-Anwendungen verwendet werden. Aufgrund der hohen Komplexität und der großen Datenmengen kann diese Architektur schnell träge werden und sich zu einem Datensumpf entwickeln.

Neben den eben aufgezeigten Veränderungen der Datenquellen, gibt es auch bei den Anforderungen der Datennutzer große Unterschiede. Während eine Abteilung Daten auf Tagesbasis benötigt, braucht eine andere Abteilung Daten in Echtzeit. Bestimmte Daten müssen aus regulatorischen Gründen dauerhaft und nachvollziehbar gespeichert werden, während andere Daten nur im Moment ihrer Erstellung von Interesse sind und keine dauerhafte Historie benötigen. Daten werden verwendet, um regulatorische Anforderungen zu erfüllen, Geschäftsentscheidungen zu treffen oder ML-Anwendungen zu trainieren. Da es schwierig ist, all diese Szenarien mit einer einzigen zentralen Datenplattform abzudecken, kann es von Vorteil sein, für verschiedene Anwendungsfälle separate Lösungen zu entwickeln, um die Effizienz zu steigern.

Eine weitere Herausforderung stellt die organisatorische Divergenz zwischen operativen und analytischen Systemen dar. Beide Welten unterliegen eigenen organisatorischen Hierarchien. Während BI-, Datenanalyse- und Data-Science- Teams unter der Leitung eines Chief Data Officers (CDO) den analytischen Bereich verwalten, wird der operative Bereich durch die einzelnen fachlichen Abteilungen und die mit ihnen zusammenarbeitenden Technologieteams verwaltet. Dies führt zu einer fragilen Integrationsstruktur. Die operativen Daten werden über ETL-Pipelines (ETL = Extract, Transform, Load) in die analytischen Systeme reingeladen. Es gibt aber i. d. R. keine Data Contracts zwischen den operativen Systemen und den ETL-Pipelines für den Zugriff auf die Daten. Daraus resultieren häufig Fehler, bei denen unvorhergesehene Änderungen im Quellsystem zu Ausfällen in den ETL-Prozessen führen.

Des Weiteren ist ein fundiertes Fachwissen über die Domänen und die Systeme unerlässlich für erfolgreiche Datenanalysen. Leider bestehen Daten-Teams oft nur aus technischen Datenexperten, die von der Fachseite zu weit entfernt sind. Um das notwendige Fachwissen zu erlangen, müssen Datenexperten mühsam Informationen von den Fachbereichen und Entwicklungsteams sammeln. Allerdings gibt es keine Anreize für diese Teams, ihr Wissen zu teilen. Hier gibt es ein systematisches Problem, das den zentralistischen Ansatz an seine Grenzen führt.[2]

Die Grundprinzipien von Data Mesh

Domain Ownership: Die Verantwortung für die analytischen Daten wird dezentralisiert auf die fachlichen Domänen übertragen, die am besten mit den Daten vertraut sind. Dies können entweder die Datenquellen oder die Datenabnehmer sein. Die Daten werden logisch je Domäne aufgeteilt und unabhängig verwaltet. Jede Domäne wird durch ein funktionsübergreifendes Domänenteam, bestehend aus Fach- und IT-Experten, eigenständig in Form eines unabhängigen IT-Systems umgesetzt. Hierdurch wird die Agilität gesteigert, da weniger teamübergreifende Abstimmungen notwendig sind. Außerdem wird die Datenqualität gesteigert, da die Datenherkunft, die Datenverarbeitung und die Datennutzung näher zusammenrücken.

[2] [Chr22] Jochen Christ, Dr. Larysa Visengeriyeva, Dr. Simon Harrer „TDWI E-Book Data Mesh" SIGS DATACOM GmbH 2022

Data as a Product: Die Verlagerung der Verantwortung zu den einzelnen Domänenteams führt zu neuen Herausforderungen, wie z. B. die Zugriffsmöglichkeit auf die Daten durch andere Domänenteams, die Kompatibilität der Daten mit Daten anderer Domänenteams und die Benutzerfreundlichkeit. Diese Herausforderungen sollen durch das Prinzip „Data as a Product" gelöst werden. Dieses Prinzip setzt voraus, dass Daten wie ein Produkt betrachtet und die Abnehmer wie Kunden behandelt werden. Damit Daten ein Produkt sind, müssen diese bestimmte Merkmale aufweisen:

- **Auffindbar**: Andere Teams sollten in der Lage sein, Datenprodukte zu finden. Dies geschieht indem die Datenprodukte mit ausreichend Metadaten angereichert werden.
- **Adressierbar:** Das Datenprodukt muss über eine adressierbare Schnittstelle verfügen, die als Einstiegspunkt verwendet werden kann, um an alle relevanten Informationen, wie z. B. die Dokumentation und die eigentlichen Daten des Datenproduktes zu gelangen.
- **Verständlich:** Datenprodukte sollten ohne Hilfe von Dritten verständlich sein. Das eigenständige Verstehen eines Datenproduktes ist ein grundlegendes Merkmal der Benutzerfreundlichkeit.
- **Vertrauenswürdig und wahrheitsgetreu:** Durch Service-Level-Objectives soll eine ausreichend hohe Datenqualität sichergestellt werden.
- **Nativ zugreifbar:** Die Daten müssen adressatengerecht abrufbar sein. Das heißt, dass die Abnehmer mit ihren gewohnten Tools (Tabellenkalkulation, SQL, Data Frames etc.) auf die Daten zugreifen können.
- **Interoperabel:** Durch Standardisierung sollen Datenprodukte mit anderen Datenprodukten kompatibel sein.
- **Eigenständig nützlich:** Ein Datenprodukt sollte einen inhärenten betriebswirtschaftlichen Wert haben, auch ohne dass die Daten erst mit anderen Datenprodukten verknüpft und korreliert werden müssen.
- **Sicher:** Datenprodukte folgen der Praxis „Security Policy als Code". Diese Policies werden zur Laufzeit von jedem einzelnen Datenprodukt angewendet. Eine Security Policy kann z. B. Berechtigungen, Aufbewahrungsfristen oder Verschlüsselungen steuern.

Ein Datenprodukt kann z. B. ein Set aus Tabellen, ein Report oder ein Dashboard sein. Wichtig anzumerken ist, dass nicht nur das Endprodukt als Datenprodukt bezeichnet wird, sondern die gesamte Peripherie, also z. B. Metadaten, Schnittstellen, Security Policies etc. ebenso fest zum Datenprodukt gehört.

Self-Serve Data Platform: Um es den einzelnen Domänenteams zu ermöglichen, Datenprodukte herzustellen und dabei die Kosten im Rahmen zu halten, ist es notwendig, eine Datenplattform mit Self-Service-Diensten bereitzustellen. Diese Datenplattform soll es autonomen Domänenteams ermöglichen, die Daten ganzheitlich zu verwalten. Dies soll dadurch realisiert werden, dass APIs (APIs = Application Programming Interfaces), Tools und Dienste zur Verfügung gestellt werden, die komplexe Technologien in der Handhabung vereinfachen und somit den Cognitive Load für die Teams, die sie nutzen, verringern. Hierdurch wird es den Domänenteams ermöglicht, die Erstellung, die Bereitstellung und die Fehlerbehebung eigenständig zu erledigen. Der Einsatz von Spezialisten verlagert sich zum vornehmlichen Einsatz von Generalisten. Lediglich der Aufbau und der Betrieb der Datenplattform erfolgt durch ein spezialisiertes Plattformteam, welches jedoch datenagnostisch operiert.

Federated Computational Governance: Um die Korrelation unabhängiger Datenprodukte zu ermöglichen, wird eine übergreifende Standardisierung benötigt. Auch ein Data Mesh benötigt eine globale Governance, die domänenübergreifend und über alle Datenprodukte hinweg gilt. Diese Governance wird gemeinsam durch Mitglieder aus allen Teams föderativ definiert und kontinuierlich an die Gegebenheiten angepasst. Dies schafft globale Anreize, die den Aufbau eines Netzes von Datenprodukten anstelle von Datensilos vorantreiben und lokale Anreize, die die Geschwindigkeit und Autonomie der einzelnen Domänen erhöhen. Außerdem werden die Policies, die durch die Governance vorgegeben werden, in die einzelnen Domänen und Datenprodukte technisch eingebettet, sodass eine föderative und computergestützte Governance entsteht.

Diese vier Prinzipien ergänzen sich gegenseitig und lösen die Herausforderungen, die sich aus ihnen ergeben. Abbildung 2 zeigt das Zusammenspiel der einzelnen Prinzipien. Das Von-Prinzip birgt Herausforderungen, die durch das Zu-Prinzip gelöst werden.

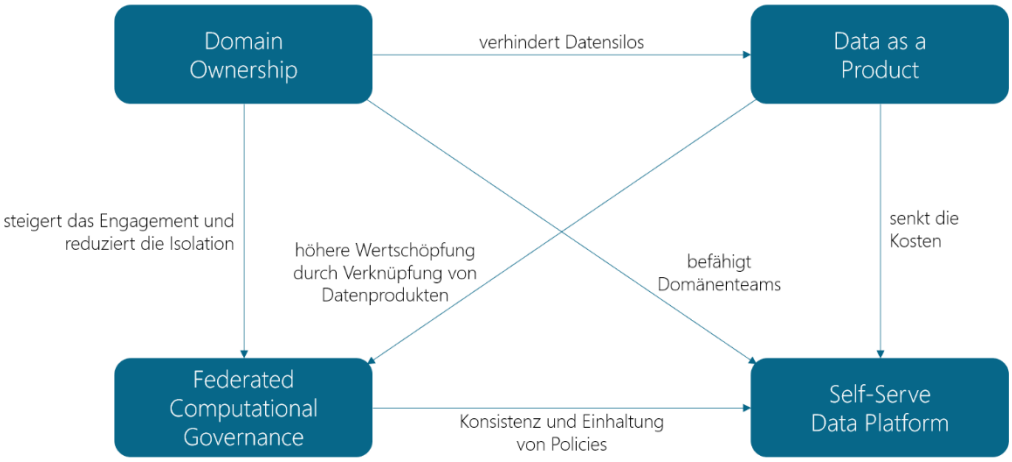

Abbildung 2: Zusammenspiel der Prinzipien[3]

Data Mesh in der Umsetzung

Da Data Mesh ein relativ neues Konzept ist, gibt es hier noch keine Best-Practice-Ansätze. Es ist hervorzuheben, dass der Fokus von Data Mesh vorwiegend auf den organisatorischen Aspekten liegt. Es ist also nicht zwingend erforderlich, direkt bei der Einführung von Data Mesh die IT-Infrastruktur großen Anpassungen zu unterziehen. Ein inkrementelles Einführungskonzept, welches zunächst ein bis zwei Domänenteams zu Beginn bildet und im Laufe der Zeit um weitere Domänenteams erweitert, könnte ein Ansatz sein. Vieles, was sich bisher bewährt hat, kann erhalten bleiben. Die Domänenorientierung beispielsweise ist bereits in vielen Datenplattformen, wie z. B. einem Data Warehouse bereits gegeben. Um sich an das Thema Data Mesh heranzutasten, ist es denkbar, dass sich zunächst mehrere Domänenteams ein bereits existierendes Data Warehouse teilen. Der Unterschied zu herkömmlichen Organisationsstrukturen sollte jedoch sein, dass jedes Domänenteam funktionsübergreifend und ganzheitlich für seine eigene Domäne zuständig ist. Außerdem sollten die Domänen keine Abhängigkeiten untereinander aufweisen, um einen klaren Fokus auf die Daten, die den Domänenteams am besten bekannt sind, zu setzen und ein autonomes Handeln zu ermöglichen.

[3] [Deh23] Zhamak Dehghani „Data Mesh – Eine dezentrale Datenarchitektur entwerfen“ O’Reilly 2023

Auch wenn der Fokus auf den organisatorischen Aspekten liegt, sollte eine mittel- und langfristige Strategie in Bezug auf die Datenmodellierung und die IT-Architektur entwickelt werden, um optimale Ergebnisse zu erzielen. Data Vault könnte ein geeigneter Modellierungsansatz für ein Data Mesh sein. Die Daten werden dabei auf Hubs, Links und Satelliten aufgeteilt. Dabei beinhalten die Hubs Schlüssel, die eine Entität eindeutig identifizieren. Dies können z. B. Konto- oder Kundennummern sein. In den Satelliten werden beschreibende Informationen gespeichert. Links bilden die Beziehungen zwischen den Entitäten ab und verbinden somit mehrere Hubs miteinander. Dies ermöglicht es, den Domänenteams an den eigenen Entitäten zu arbeiten und diese über Links mit Entitäten anderer Domänenteams zu verknüpfen. Auch die Konstellation, dass mehrere Domänenteams an der gleichen Entität arbeiten, ist umsetzbar, da die beschreibenden Attribute zu einer Entität auf mehrere Satelliten aufgeteilt werden können, sodass jedes Domänenteam an seinem eigenen Satelliten arbeiten kann, wie beispielhaft in Abbildung 3 dargestellt. [4]

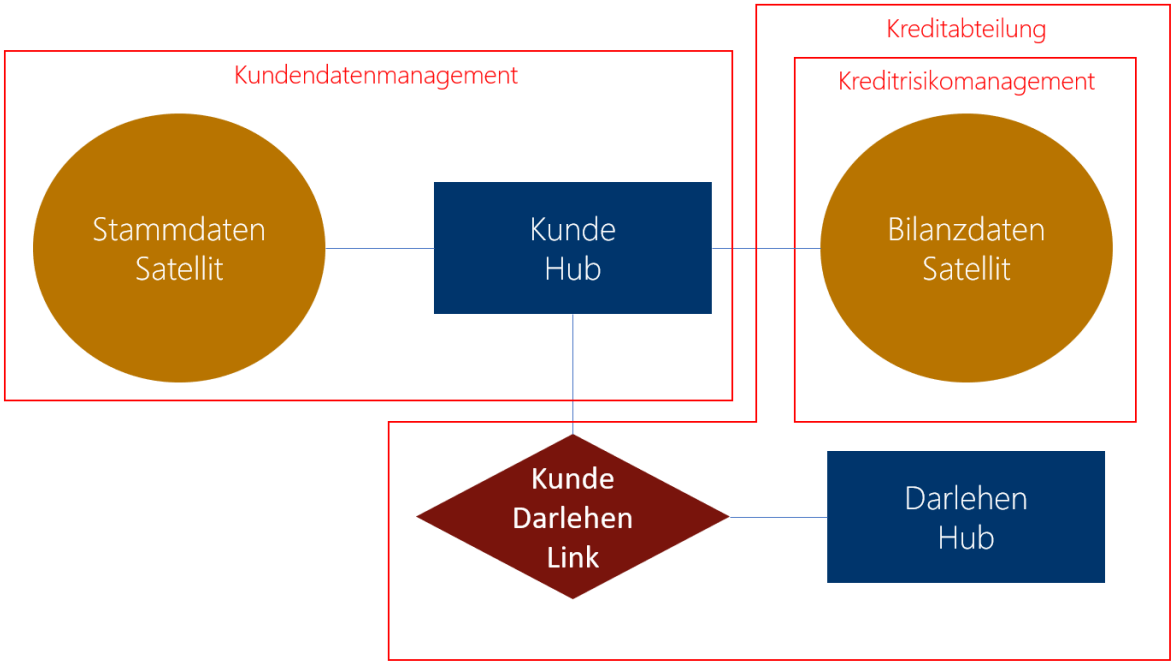

Abbildung 3: Beispiel Data-Vault-Modellierung

Der hohe Standardisierungsgrad und die klaren Modellierungskonventionen von Data Vault bieten ein hohes Automatisierungspotential. Mittlerweile gibt es zahlreiche Generatoren für Data Vault auf dem Markt (z. B. VaultSpeed, Datavault Builder, WhereScape etc.). Der Einsatz eines Generators hilft dabei die Aufwände und das erforderliche technische Know-how bei der Implementierung zu reduzieren. Wenn Sie mehr über Data Vault erfahren möchten, lesen Sie auch unseren Artikel **Data Warehousing & Data Vault 2.0**.

[4] [Kau22] Christian Kaul „Warum Data Vault und neue Datenarchitekturansätze zusammenpassen" BI-Spektrum 04/2022

Um eine Data-Mesh-Architektur, wie in Abbildung 4 dargestellt, umzusetzen, können die Domänenteams sowohl auf der gleichen Datenbank, auf separaten Datenbanken mit gleicher Technologie (z. B. Snowflake, Azure, Oracle etc.) oder auch auf verschiedenen Datenbanken mit unterschiedlicher Technologie arbeiten. Separate Datenbanken mit gleicher Technologie werden dabei z. B. über Datenbanklinks miteinander verbunden und Datenbanken mit unterschiedlichen Technologien können über Datenvirtualisierungslösungen, wie z. B. Denodo, Stone Bond oder Data Virtuality zusammengeführt werden. Datenvirtualisierungsplattformen bieten häufig integrierte Governancelösungen an, die es ermöglichen, systemübergreifende Policies z. B. für Datensicherungen, Datenzugriffe und Datenanonymisierung zu implementieren. Mit der Google Cloud gibt es bereits einen der ersten Ansätze für ein umfangreiches Toolkit, das es ermöglicht eine Self-Serve Data Platform bereitzustellen, mit der Domänenteams in der Lage sein sollen, Datenprodukte eigenständig zu entwickeln und bereitzustellen. [5]

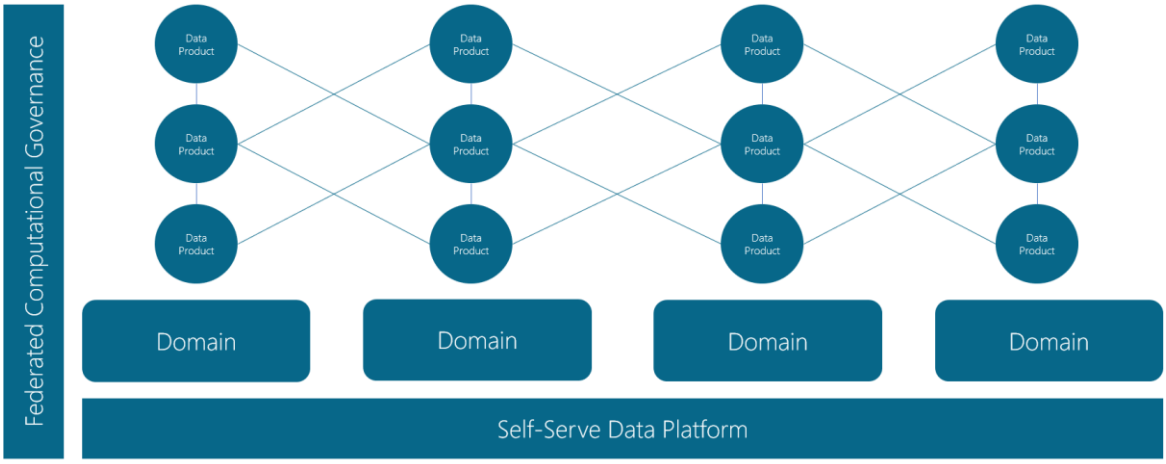

Abbildung 4: Logische Architektur von Data Mesh

Auch Themen wie Testautomatisierung tragen dazu bei, dass die Domänenteams entlastet werden und Datenprodukte schneller bereitgestellt werden können. Tools wie z. B. Tricentis Tosca ermöglichen es, automatisierte Testfälle auch ohne Programmierkenntnisse zu erstellen. Wenn Sie mehr über Tricentis Tosca erfahren möchten, lesen Sie auch unseren Artikel **SMARTER mit Tricentis Tosca®**.

[5] [Tek22] Firat Tekiner, Thinh Ha, Johan Picard, Victor Crowther, Susan Pierce „White paper – Build a modern distributed Data Mesh with Google Cloud" Google 2022

Fazit

Data Mesh kann als eine Sammlung von bereits etablierten Konzepten (Domain Driven Design, Product Thinking, Platform Thinking etc.) betrachtet werden, durch deren Kombination ein eigenständiges Konzept entstanden ist. Wenn man den Wandel betrachtet, der durch die Einführung agiler Methoden in der Unternehmenskultur entstanden ist, dann stellt man fest, dass Data Mesh die logische Konsequenz daraus ist. So findet sich beispielsweise der Einsatz von autonomen und interdisziplinären Teams als Kerndisziplin auch in Data Mesh wieder. Häufig werden viele Aspekte, die Data Mesh auszeichnen, in Projekten bei der Einführung von neuen Systemen bereits gelebt. Wenn der Wechsel vom Projektmodus in den Betriebsmodus stattfindet, fallen allerdings viele Unternehmen in ihre alten organisatorischen Muster zurück. Um das Ziel zu erreichen, eine Data Driven Company zu sein, müssen die organisatorischen Strukturen so angepasst werden, dass diese flexibel und skalierbar sind. Genau dieses Potential birgt Data Mesh.

Zu den Kernherausforderungen bei der Einführung von Data Mesh gehören die Zusammenstellung und Einführung eines auf das Unternehmen abgestimmten Technologiestacks, das Self-Service-Dienste für die Domänenteams bereitstellt und die Etablierung einer neuen Datenkultur. Datenproduzenten und -konsumenten müssen zusammengebracht werden, um den Austausch von Perspektiven und Zielen zu ermöglichen. Technologien können die Zusammenarbeit bei der Datenanalyse effektiv unterstützen. Letztendlich sind es jedoch die User, die durch ihre Beiträge die Technologie zum Leben erwecken.

Wie Finbridge Sie unterstützt

Finbridge unterstützt Sie bei der Entwicklung Ihrer zukunftssicheren Datenstrategie und der entsprechenden IT-Architektur. Durch unsere umfassende technische Expertise und unser fachliches Know-how in der Banken- und Finanzdienstleistungsbranche können wir flexibel auf die spezifischen Bedürfnisse Ihres Unternehmens eingehen und Sie bei der Umsetzung einer zukunftsorientierten IT- und Datenlandschaft begleiten, die auch allen regulatorischen Anforderungen gerecht wird.

Quellen

[Deh23] Zhamak Dehghani „Data Mesh – Eine dezentrale Datenarchitektur entwerfen“ O’Reilly 2023

[Chr22] Jochen Christ, Dr. Larysa Visengeriyeva, Dr. Simon Harrer „TDWI E-Book – Data Mesh“ SIGS DATACOM GmbH 2022

[Kau22] Christian Kaul „Warum Data Vault und neue Datenarchitekturansätze zusammenpassen“ in BI-Spektrum 04/2022 SIGS DATACOM GmbH 2022

[Tek22] Firat Tekiner, Thinh Ha, Johan Picard, Victor Crowther, Susan Pierce „White paper – Build a modern distributed Data Mesh with Google Cloud“ Alphabet Inc. 2022

Autor

Ilja Jost

Expert
Solutions
LinkedIn

FINBRIDGE

Insights und Trends

Finbridge GmbH & Co. KG
Louisenstraße 100
61348 Bad Homburg v. d. H.
www.finbridge.de