Analytische Datenarchitekturen im Wandel: Eine Analyse von Data Mesh und Data Lake

Ibrahim Al-Showiter

Technische Hochschule Mittelhessen

Mathematik, Naturwissenschaften und Datenverarbeitung
Wilhelm-Leuschner-Str. 13
61169 Friedberg
E-Mail:
ibrahim.ali.mohammed.ahmed.al-showiter@mnd.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule Mittelhessen

Mathematik, Naturwissenschaften und Informatik
Wiesenstraße 14
35390 Gießen
E-Mail:
harald.ritz@mni.thm.de

Prof. Dr. Frank Kammer

Technische Hochschule Mittelhessen

Mathematik, Naturwissenschaften und Informatik
Wiesenstraße 14
35390 Gießen
E-Mail:
frank.kammer@mni.thm.de

Kategorie

Bachelorarbeit

Schlüsselwörter

Big Data, Business Intelligence, Data Lake, Data Mesh, (Analytische-) Datenarchitekturen

Zusammenfassung

Im Zeitalter der Informationsflut ist ein enormes Wachstum der Datenmenge und -komplexität zu verzeichnen. Dies hat die Notwendigkeit fortschrittlicher Technologien und Strategien zur Datenverwaltung in den Vordergrund gerückt. In diesem Kontext stechen zwei Lösungsansätze besonders hervor: Data Lake und Data Mesh. Diese beiden Ansätze stehen im Mittelpunkt der vorliegenden Bachelorarbeit und repräsentieren unterschiedliche Herangehensweisen in der Datenarchitektur, die für Organisationen von großer Bedeutung sind.

Die Hauptzielsetzung dieser Bachelorarbeit besteht darin, Data Mesh und Data Lake durch eine gründliche Analyse ihrer zentralen Charakteristika und Unterschiede miteinander zu vergleichen. Dabei liegt der Fokus auf der Untersuchung der Hauptelemente beider Datenarchitekturen sowie ihrer Relevanz in der modernen Datenverwaltung. Die Arbeit beabsichtigt auch zu erörtern, inwieweit diese Ansätze unterschiedliche oder möglicherweise ergänzende Anforderungen erfüllen können.

Im Rahmen dieser Untersuchung wurden folgende zentrale Fragen behandelt:

1. Welche spezifischen Merkmale zeichnen die Konzepte von Data Lake und Data Mesh aus, und welche wesentlichen Unterschiede bestehen zwischen ihnen?

2. Unter welchen Bedingungen sollten Organisationen, die Data Lake-Architekturen nutzen, über einen möglichen Übergang zu Data Mesh nachdenken?

Diese Fragen bildten das Grundgerüst der Untersuchung und ermöglichten eine tiefgehende Analyse der Datenarchitekturen Data Lake und Data Mesh im Hinblick auf ihre Charakteristika und potenziellen Anwendungsgebiete.

Die Untersuchung beginnt mit einer einführenden Darstellung der Entwicklung und Evolution hin zu den aktuellen analytischen Datenarchitekturen Data Lake und Data Mesh. Dieser Kontext legt den Grundstein für die folgende eingehende Analyse der beiden Ansätze.

Im Hauptteil der Arbeit steht die detaillierte Untersuchung und Analyse der Data Lake- und Data Mesh-Architekturen im Fokus. Dieser Abschnitt beginnt mit der Explorierung ihrer zugrundeliegenden Prinzipien sowie ihres jeweiligen grundlegenden Aufbau. Dabei wurde die Analyse von den unterschiedlichen Architekturen von Data Lake an dem im Zeitschriftenaufsatz „On data lake architectures and metadata management“ vorgestellten Ansatz orientiert. Der Ansatz heißt „Functional × Maturity“ und dies ermöglicht eine präzisere Untersuchung von Data Lake Architekturen, indem er zwischen funktionalen, auf dem Reifegrad der Daten basierenden sog. Data Maturity und hybriden Architekturen unterscheidet.

Bei der Untersuchung von der Architektur und Prinzipien von Data Mesh wurden die vier Prinzipien nämlich: Domain Ownership, Data as a Product, Self-Serve Dataplattform, Federated Computational Governance näher betrachtet.

Anschließend wird die Implementierung beiden Ansätze betrachtet. Dabei werden bei Data Lake verschiedene Technologien und Ansätze vorgestellt, die bei der Erstellung eines Data Lakes zum Einsatz kommen können. Die Umsetzung von Data Mesh berücksichtigt entscheidende Aspekte. Dies beinhaltet die Implementierung domänenspezifischer Schnittstellen für Datenanalysen sowie das zentrale Konzept von Datenprodukten als Architekturquanten. Hierbei werden strukturelle Komponenten wie Code, Daten, Metadaten und Plattformabhängigkeiten berücksichtigt. Ein weiterer Fokus liegt auf der effektiven Bereitstellung und dem Konsum von Daten, ergänzt durch Discovery- und Observability-APIs.

Nachfolgend werden verschiedene Aspekte wie die Implementierung, Verwaltung und organisatorische Einbettung beider Ansätze behandelt. Dieses Kapitel schließt mit einer kritischen Auseinandersetzung der Herausforderungen und Hürden ab, die bei der Anwendung von Data Lake und Data Mesh auftreten können.

Ein nachfolgendes Kapitel widmet sich dem Vergleich der beiden Ansätze, wobei spezifische Aspekte wie der organisatorische und technologische Rahmen, die Zusammenarbeit und Governance, Datenanalyse sowie die Anwendbarkeit im Bereich Machine Learning (ML) und Künstliche Intelligenz (KI) betrachtet werden. Dieser systematische Vergleich ermöglicht einen umfassenden Einblick in die Unterschiede und Gemeinsamkeiten beider Ansätze.

Im letzten Kapitel der Arbeit wurden die gewonnenen Erkenntnisse zusammengeführt und Schlussfolgerungen gezogen. Es wird deutlich, dass Data Lakes vorrangig als zentrale Speicherlösungen konzipiert sind und ihren Schwerpunkt auf technologischen Aspekten legen. Im Gegensatz dazu verfolgt Data Mesh einen Ansatz der organisatorischen Neuausrichtung, bei dem Daten dezentralisiert und als eigenständige Produkte betrachtet werden. Diese Unterscheidung bildet die zentrale Erkenntnis, die sich als Antwort auf die zweite Fragestellung herauskristallisiert: Ein Übergang zu Data Mesh sollte in Betracht gezogen werden, wenn Organisationen mit Skalierungsproblemen in ihrer aktuellen Data Lake-Architektur konfrontiert sind.

Literatur

Butte, Vijay Kumar; Butte, Sujata (2022): Enterprise Data Strategy: A Decentralized Data Mesh Approach.

Dehghani, Zhamak (2023): Data Mesh. Eine dezentrale Datenarchitektur entwerfen. 1. Auflage. Heidelberg: O'Reilly.

Fang, Huang (2015): Managing data lakes in big data era: What's a data lake and why has it became popular in data management ecosystem.

Gartner (2014): Gartner Says Beware of the Data Lake Fallacy. Online verfügbar unter https://www.gartner.com/en/newsroom/press-releases/2014-07-28-gartner-says-beware-of-the-data-lake-fallacy

Hlupic, Tomislav; Orescanin, Drazen; Ruzak, Domagoj; Baranovic, Mirta (2022): An Overview of Current Data Lake Architecture Models. In: Neven Vrcek (Hg.): 2022 45th Jubilee International Convention on Information, Communication and Electronic Tech-nology (MIPRO).

Khine, Pwint Phyu; Wang, Zhao Shun (2018): Data lake: a new ideology in big data era.

Mehmood, Hassan; Gilman, Ekaterina; Cortes, Marta; Kostakos, Panos; Byrne, Andrew; Valta, Katerina et al. (2019): Implementing Big Data Lake for Heterogeneous Data Sources. In: 2019 IEEE 35th International Conference on Data Engineering work-shops.

Sharma, Ben (2018): Architecturing Data Lakes - Zaloni. Data Management Architectures for Advanced Business Use Cases. Zaloni.

Papp, Stefan; Weidinger, Wolfgang; Meir-Huber, Mario; Ortner, Bernhard; Langs, Georg; Wazir, Rania (2019): Handbuch Data Science. Mit Datenanalyse und Machine Learning Wert aus Daten generieren. München: Hanser (Hanser eLibrary).