# Data-Lake-Governance-Konzepte

Sohrab Sahba

Technische Hochschule
Mittelhessen

Fachbereich Mathematik,
Naturwissenschaften und
Datenverarbeitung
Wilhelm-Leuschner-Straße 13
61169 Friedberg
sohrab.sahba@mnd.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule
Mittelhessen

Fachbereich Mathematik,
Naturwissenschaften und
Informatik
Wiesenstraße 14
35390 Gießen
harald.ritz@mni.thm.de

**Kategorie**

Bachelorarbeit

**Schlüsselwörter**

Data Lake, Data Governance, Data Lake Governance, Data Steward, Data-Lifecycle-Management, Datensicherheit, Data Lineage, Datenqualitätsmanagement, Metadatenmanagement, Schema-on-Read, Schema-on-Write, Repository

**Zusammenfassung**

Im Zeitalter der Digitalisierung werden täglich eine Vielzahl von Daten generiert. Hierdurch beschäftigen sich Organisationen mehrheitlich mit dem schnellen Wachstum der Datenmengen und inwiefern ein Mehrwert zur Steuerung des Unternehmens erfolgen kann. Um dies zu ermöglichen, müssen Daten bedarfsgerecht gespeichert und gepflegt werden. Zur Verständlichkeit, Wiederverwendbarkeit und Vertrauenswürdigkeit der Daten in Informationssystemen ist die Einführung einer Data Governance erforderlich.

Durch die Generierung von unterschiedlichen Datenarten, wie Bilder bzw. Daten aus dem Social-Media-Bereich oder E-Mails, wird ein Repository benötigt, welcher im Gegensatz zu klassischen Data-Warehouse-Systemen diese unterstützt. Eine Möglichkeit zur Datenhaltung bildet der Data Lake. Er bietet neben strukturierten Daten auch die Verwaltung und Speicherung von semi- und unstrukturierten Daten an. Der Data Lake verwendet das Schema-on-Read Verfahren, welches Daten aus der Datenquelle extrahiert und anschließend in das Zielsystem speichert. Sobald die Daten für eine Analyse benötigt werden, können diese zentral aus dem Data Lake abgefragt werden.

Bei Data Governance handelt es sich um die Zuweisung von Entscheidungsrechten und den damit verbundenen Pflichten bei der Verwaltung von Daten. Data Governance stellt organisationsweit eine Sekundärorganisation dar, welche den Wert der Daten maximieren soll. Durch die Einführung von Data Stewardship können bereichsübergreifend Verantwortlichkeiten zugeordnet werden, damit Data Stewards sich um die Daten der Organisation kümmern können.

Das Ziel der Arbeit liegt darin, anhand der aktuell verfügbaren Literatur ein Data-Lake-Governance-Konzept zu erstellen, das universell anpassbar ist. Es wird eine Relation zwischen dem Data Lake und der Data Governance geschaffen, indem Rollen und Handlungsfelder der Data Governance in Bezug zum Data Lake vorgestellt und definiert werden. Aus diesen Rollen und Handlungsfeldern ergeben sich anschließend Zuständigkeiten zur Etablierung der Data Lake Governance.

Anhand von cloud-basierten Lösungen aus dem Hause Microsoft, wurde mit Azure ein Data-Lake-Ökosystem implementiert. Die seitens Microsoft Azure verfügbaren Komponenten, wie Azure Purview, Azure Data Lake Storage Gen2 und Azure Data Factory ermöglichen es, die zuvor definierten Methoden der Data Lake Governance ohne großen Aufwand einzuführen und anzuwenden.

Das Ziel der Data Lake Governance ist es, die Entwicklung eines Data Lake in einen Data Swamp zu vermeiden. Durch die Einführung eines Data-Lake-Governance-Rahmenwerks ist die bedarfsgerechte Verwaltung und Speicherung von Daten garantiert, wodurch jeglichen Anwendern ein einheitliches Verständnis der Daten zugesichert wird.