Metadatenmanagement bei Data-Lake-Realisierungsansätzen

Dogus Tansel

Technische Hochschule Mittelhessen

Fachbereich MND
Wilhelm-Leuschner-Straße 13
61169 Friedberg
E-Mail: dogus.tansel@mnd.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule Mittelhessen

Fachbereich MNI
Wiesenstraße 14
35390 Gießen
E-Mail: harald.ritz@mni.thm.de

Kategorie

Bachelorarbeit

Schlüsselwörter

Metadatenmanagement, Metadaten, MDMS, Data Lake, Data Factory, Metadaten-Repository, Metadatenpflege, Schema-on-Read

Zusammenfassung

Im Zeitalter von Big Data werden täglich mehrere Billionen Daten weltweit erzeugt und zwischen unzähligen Systemen bewegt und verarbeitet. Sei es von industriellen Maschinen, AI-Software oder vom Menschen selbst erschaffen, so müssen diese enormen Datensätze gespeichert und gepflegt werden. Im Zuge der Digitalisierung ist es für Unternehmen umso wichtiger ausreichende Informationen über ihre Daten in Form von Metadaten zu erhalten, um Transparenz zu erzeugen, wie diese aufgebaut und verwendet werden. Diese gewaltigen Mengen an divergierenden Daten müssen in Systemen auf sorgfältiger Weise abgelegt und verwaltet werden, um einen ausreichenden Mehrwert aus den gewonnen Informationen ziehen zu können.

Heutzutage fallen eine Vielfalt von möglichen Arten von Daten an, welche nicht mehr in klassischen Datenbanken bzw. Data-Warehouse-Systemen für Anwendungsfälle im Bereich Business Intelligence und Data Analytics gespeichert werden können, wie z.B. Bilder, Musik, Serverdaten u.v.m. Um solch ein breites Cluster an Daten abfangen und im jeweiligen Fall nutzen, sowie mithilfe von Metadaten in einem verwertbaren Rahmen verwalten zu können, bedarf es einer neuen und zeitgetreuen Alternative zu herkömmlichen Speicherverfahren. Ein Data Lake bietet diese Möglichkeit zum Verwalten und Analysieren heterogener, komplexer sowie umfangreicher Daten jeglicher Art, um die Realisierung eines Metadatenmanagementansatzes anzutreiben. Für dieses Unterfangen ist das Schema-on-Read-Verfahren in einem Data Lake zu realisieren. In einem klassischen Data Warehouse bildet die verwendete Datengrundlage ein Konstrunkt an Informationen, das bereits zu Beginn des Ladeprozesses transformiert worden ist. Mit dem Schema-on-Read-Verfahren werden die verwendeten Daten erst zu Beginn einer Notwendigkeit in ihre entsprechende Form transformiert und somit dem Anwender innerhalb des Data Lakes zur Verfügung gestellt.

Mit der vorliegenden Abschlussarbeit wird der Frage nachgegangen, welche Rolle das Metadatenmanagement in der heutigen Zeit einnimmt und wie dieses in einem modernen und zeitgetreuen Datenbankansatz zu realisieren ist. Hinsichtlich dieses Themas werden Implementierungsbeispiele für das Metadatenmanagement in Data Lakes mit strukturierten, semi-strukturierten und sog. unstrukturierten Daten realisiert, die Bedeutung von Metadatenmanagement-Konzepten und deren Technologien nähergebracht sowie die positiven und negativen Aspekte eines solchen Ansatzes dargestellt.

Als Ergebnis wurde deutlich, dass durch das Enststehen von Metadaten Unternehmen dazu getrieben werden ihre bisherigen Kenntnisse des Datanmanagements zu erweitern und sich im Spektrum einer metadatengetriebenen Architektur des Metadatenmanagements einzuordnen haben, um den Nutzern einer internen BI-Plattform im Rahmen von Self-Service- und Governance-Anforderungen gerecht zu werden.

Weiterhin ist mit der Implementierung eines Data-Lake- und Metadatenmanagement-Systems mit cloudbasierten Komponenten des Anbieters Azure und Google Cloud Products festzuhalten, dass das Metadatenmanagement eine zu bewältigende Herausforderung für Unternehmen und Industrie darstellt, um konsequente Datenhaltung über Systemgrenzen hinweg zu gewährleisten. Durch den Einsatz von ausgewählten IT-Werkzeugen, wie z.B. der Data Factory, die als cloudbasierter Datenintegrationsdienst die Pflege und das Verwalten großer Mengen von Informationen, insbesondere von Metadaten, ermöglicht, ist das Erfüllen der zentralen Ziele des Metadatenmanagements gewährleistet.

Diese Ziele sind zum einen die Minimierung der Aufwände beim Erstellen und im Betrieb von Informationssystemen mit großen Mengen an Daten und Metadaten und zum anderen das Erreichen eines weitreichenden Mehrwerts in der Informationsgewinnung für Entwickler und Endbenutzer.