Datenmodellierungsansätze für Data Lakes

Liam-Theodor Heidemann

Technische Hochschule Mittelhessen

Fachbereich MNI
Wiesenstr. 14
35390 Gießen
E-Mail: liam-theodor.heidemann@mni.thm.de

Prof. Dr. Harald Ritz

Technische Hochschule Mittelhessen

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

.

Kategorie

Bachelorarbeit

Schlüsselwörter

Data Lakes, Datenmodellierung, Datenmodellierung-sansätze, Big Data Schema-on-Read

Zusammenfassung

Durch die Weiterentwicklung des Internets und der damit einhergehenden Digitalisierung, werden weltweit immer mehr Daten generiert. Im Bereich Big Data wird ein Wachstum der Datenmenge des Jahres 2016 auf über das achtfache im Jahr 2021 prognostiziert. Aus diesen enormen Datenmengen entstehen Herausforderungen für Unternehmen wie das Speichern der Datenmengen, die Komplexität der Analyse, das zusammenwachsen von strukturierten und unstrukturierten Daten uvm.

Zur Bewältigung dieser Herausforderungen wurde im Jahr 2010 das Konzept des Data Lake vorgestellt. Ein Data Lake verfolgt das Ziel, alle Daten in ihrer Rohform an einem zentralen Punkt kostengünstig abzuspeichern. Die Abfrage der Daten findet nach dem Konzept „Schema-on-Read“ statt, sodass die Daten erst bei der Nutzung in das gewünschte Schema gebracht werden.

Im Laufe der Zeit hat sich gezeigt, dass ein Data Lake schnell zu einem Data Swamp werden kann, wenn alle Daten ohne Kontrolle und grundlegendes Datenmodell abgespeichert werden. So entstehen ohne Datenmodell Probleme hinsichtlich der Datenqualität, der Daten-verständlichkeit und der Informationsintegration.

Das Ziel der Arbeit lag darin, auf Grundlage der aktuell verfügbaren Literatur den Ansatz des Data Lake vorzustellen, einen Überblick zum Thema Daten-modellierung zu geben, die beiden Themen-bereiche zusammenzuführen und abschließend bereits bestehende Datenmodellierungsansätze für Data Lakes vorzustellen, zu vergleichen und zu evaluieren.

In diesem Zuge wurde der Begriff des Data Lake umfassend erläutert und vom Data Warehouse abgegrenzt, es wurde die Motivation zum Einsatz eines Data Lakes dargestellt, verschiedene positive wie auch negative Stimmen betrachtet und der mögliche Aufbau eines Data Lake anhand zweier Referenzarchitekturen vorgestellt.

Zum Thema der Datenmodellierung wurde der Begriff in seinen Ausprägungen definiert, die Ziele, die die Datenmodellierung verfolgt dargestellt und sowohl die relationale wie auch die multidimensionale Datenmodellierung vorgestellt. Abschließend wurden die Themen Datenmodellierung und Data Lakes zusammen-geführt, drei Ansätze zur Modellierung vorgestellt, welche anhand von Vergleichskriterien betrachtet und im Ergebnis evaluiert wurden.

Das Ergebnis der Arbeit zeigt, dass von den vorgestellten Ansätzen der Data-Droplet-Methode, der Fragment-Repräsentation und Data Vault bisher lediglich Data Vault für den Umgang mit strukturierten Daten im Data Lake umgesetzt wurde. Die anderen Ansätze sind zum jetzigen Stand nur in der Theorie oder in Modellversuchen umgesetzt worden, bieten allerdings Potenzial, weiter erforscht zu werden.

Es ist zu beachten, dass die Arbeit sich nur mit aktuell in der Literatur verfügbaren Quellen beschäftigt hat. Ggf. gibt es bereits in der Wirtschaft weitere Erkenntnisse zur Umsetzung der Datenmodellierung innerhalb eines Data Lake.