B. Mitschang et al. (Hrsg.): Datenbanksysteme für Business, Technologie und Web (BTW 2017),
Lecture Notes in Informatics (LNI), Gesellschaft für Informatik, Bonn 2017 527

Drying up the data swamp – Vernetzung von Daten mittels iQser GIN Server

Florian Pfleiderer[1]

Abstract

In vielen Unternehmen laufen heute heterogene Daten aus vielfältigen Quellen in Data Lakes zusammen, die immer mehr zu Data Swamps verkommen. Oft ist nicht bekannt, was sich in den zahlreichen Datentöpfen befindet und in welcher Qualität die Daten tatsächlich vorliegen. Typische Big Data Technologien wie zum Beispiel Hadoop alleine bieten kaum eine Möglichkeit, diesem Chaos Herr zu werden. Immer mehr Firmen zeigen daher Interesse an kompletten Lösungen, statt eigene Lösungen aufwändig aus einzelnen Technologien zusammen zu stellen. Die iQser GmbH entwickelt mit dem GIN Server eine solche Lösung, die unterschiedliche Ansätze des Data Engineering kombiniert, um verschiedenste Problemstellungen der semantischen Datenanalyse lösen zu können.

Um aus einem Mix von strukturierten und unstrukturierten Daten Informationen gewinnen zu können, werden Daten und Dokumente basierend auf ihren Inhalten mithilfe qualifizierter Relationen automatisch vernetzt. Der hierbei entstehende Graph ist die Basis für die Schöpfung von neuem Wissen aus vorher unbekannten Daten. Solche Daten können nicht immer im Vorfeld klassifiziert oder auf bestimmte Arten modelliert werden, da hierfür das notwendige a-priori Wissen über die Inhalte der Daten fehlt oder zu aufwändig zu erlangen ist. Dies betrifft insbesondere die Erstellung von Ontologien im Sinne des Semantic Web oder Open Linked Data. Hier geht die Lösung von iQser einen anderen Weg und erzeugt in einem Bottom-Up-Ansatz aus den Daten selbst ein Modell über eine automatische semantische Vernetzung.

In dem Vortrag wird erklärt, welche Ziele mit der Entwicklung des GIN Servers verfolgt wurden, um Ordnung in einem Data Swamp zu schaffen, in dem mehr Daten nicht immer mehr Nutzen bedeuten, weil es immer schwerer wird diese zu korrelieren und ordnen zu können. Es wird darauf eingegangen, welchen Herausforderungen man sich bei der Entwicklung einer solchen Lösung stellen muss, welche Erfahrungen gemacht und Erkenntnisse hierbei gewonnen wurden und warum ein Schritt weg von einer Batch-Verarbeitung und hin zu einem Streaming-basierten Ansatz es der Anwendungsarchitektur ermöglicht hat, Ziele besser zu erreichen.

[1] dibuco GmbH, Franz-Schubert Str. 75, 70195 Stuttgart, florian.pfleiderer@dibuco.de