Datengetriebener Schema-Evolutions-Prozess zur Erkennung von Multitype-Schema-Operationen in NoSQL-Datenbanken

Dominik Ludwig

Hochschule Darmstadt

Betreuer(in): Prof. Dr. Uta Störl und Prof. Dr. Peter Muth

Fachbereich: Fachbereiche Mathematik und Naturwissenschaften & Informatik

Motivation und Forschungsfragen

Die Prinzipien der agilen Arbeitsweise, kurze Entwicklungszyklen, Kundennähe und Kundenverständnis sind in der Datenmodellierung angekommen. Änderungen am Datenmodell sollen zu jedem Zeitpunkt im Entwicklungszyklus möglich sein. Haupttreiber solcher Änderungen sind Anwendungen, die Daten konsumieren. Eine Veränderung eines Datenmodells beeinflusst Datenkonsumenten. Die Entwicklung eines Datenmodells findet unabhängig von der Anwendungsentwicklung statt. Dies wird durch eine Data-Lake-Architektur auf die Spitze getrieben. Der Datenbankentwickler auf der einen Seite arbeitet mit vielen Änderungen am Datenmodell. Auf der anderen Seite erwartet der Anwendungsentwickler eine stabile Schnittstelle zu denselben Daten. Das führt zu einem Spannungsverhältnis, weil Änderungen des Datenmodells, zumal in hoher Frequenz, nicht dazu führen können, dass die Schnittstelle ständig unerwartete Ergebnisse liefert. Um dieses Spannungsverhältnis zu überbrücken, wird ein Schema-Evolutions-Prozess eingeführt.

Evolution ist ein kontinuierlicher Verbesserungsprozess sowie die Weitergabe vorteilhafter Eigenschaften an eine Kindgeneration eines Lebewesens. Analog dazu werden in der Schema-Evolution vorteilhafte Eigenschaften an spätere Versionen desselben Datenmodells weitervererbt. Schema-Evolutions-Prozesse begleiten die Schema-Evolution dergestalt, dass Änderungen am Schema den Konsumenten der Daten in einem geordneten Verfahren bekannt gemacht werden.

Multitype-Schema-Operationen sind entitätsübergreifende Schema-Operationen. Solche werden durch ein Kopieren oder Verschieben von Attributen verursacht. Datenkonsumenten sind an Multitype-Schema-Operationen interessiert, da solche Schema-Veränderungen eine semantische Veränderung einer bestehenden Schnittstelle darstellen.

In Zusammenarbeit mit der Commerzbank AG ist ein datengetriebener Schema-Evolutions-Prozess für eine Data-Lake-Architektur unter der Verwendung der NoSQL-Datenbank Apache Hive prototypisch entwickelt worden. Durch eine kontinuierliche Datenanalyse können Schema-Veränderungen erkannt werden. Das Thema Schema-Evolution ist kein theoretisches Konstrukt. Innerhalb eines Monats werden bei der Commerzbank in etwa 2.500 Schema-Veränderungen registriert. Diese hohe Anzahl motiviert die Notwendigkeit nach einem automatisierten Schema-Evolutions-Prozess. Die Datenanalyse umfasst die folgenden Schritte: Ähnlichkeitsberechnung, Verbundberechnung und Schema-Operations-Extraktion. Mithilfe von Ähnlichkeiten werden Attribute identifiziert, die durch einen Evolutionsprozess betroffen sind. Durch eine Verbundberechnung wird zwischen diesen Attributen eine Zeilenbeziehung hergestellt. Abschließend können Multitype-Schema-Operationen aus der Schema-Historie gewonnen werden.

Daten

Der TPC-H-Benchmark entstammt einer Familie von Benchmarks, die unterschiedliche Einsatzszenarien für Datenbanken modellieren und evaluieren [1][vgl. 12]. Dieses Datenmodell wird für die Evaluation des datengetriebenen Schema-Evolutions-Prozess zur Erkennung von Multitype-Schema-Operationen verwendet, da ein realitätsnahes Datenmodell beschrieben wird. Neben dem TPC-H-Datenmodell werden Datenmodelle aus dem Kontext der Commerzbank verwendet. Multitype-Schema-Veränderungen werden manuell in die Datenmodelle eingebracht.

Für eine breite Abdeckung von NoSQL-Datenbanken werden keine Schema-Informationen verwendet, die durch eine NoSQL-Datenbank bereitgestellt werden. Stattdessen werden Schemata datengetrieben erhoben. So kann eine breite Abdeckung von NoSQL-Datenbanken erreicht werden.

Schema-Evolutions-Prozess

Mithilfe eines Schema-Evolutions-Prozesses werden Veränderungen an Datenmodellen offengelegt. Datenmodellen widerfährt ein kontinuierlicher Veränderungsprozess, den Datenbankentwickler, zum Beispiel durch veränderte Anforderungen, verursachen [6][vgl. 2764].Eine Schema-Veränderung wird durch eine Schema-Operation eingeführt. Es können zwei Schema-Operationstypen unterschieden werden [6][vgl. 2767]:

Singletype-Schema-Operation: Diese Schema-Operationen wirken sich auf genau eine Entität aus. Sie umfassen die Operationen *Add*, *Remove* und *Rename*.

Multitype-Schema-Operation: Es sind mehrere Entitäten an einer Schema-Operation beteiligt. Vertreter dieser Gruppe sind *Copy* und *Move*. Zusätzlich wird ein Verbund der Entitäten benötigt. Ein Verbund zweier Entitäten beschreibt, wie eine Entität zu einer anderen Entität in Relation gebracht werden kann.

Durch eine alleinige Betrachtung von Schema-Informationen können Multitype-Schema-Operationen nur teilweise erkannt werden. Das folgende Beispiel zeigt die Grenze dieses Verfahrens und motiviert die Notwendigkeit eines datengetriebenen Schema-Evolutions-Prozesses.

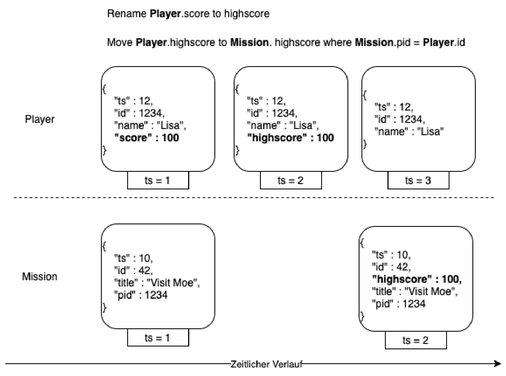


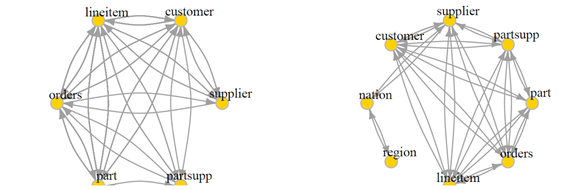
Figure: Move-Schema-Operation nach [5][vgl. 67]

Das Attribut *score* wird der *Player*-Entität entnommen und der *Mission*-Entität mit dem geänderten Namen *highscore* hinzugefügt. Durch eine alleinige Betrachtung der Schema-Veränderungen kann nicht entschieden werden, welche Schema-Operation durchgeführt wurde. Die Attribute *score* und *highscore* sind lexikalisch verschieden. Die Schema-Veränderungen können nicht zueinander in Beziehung besetzt werden. Mithilfe eines datengetrieben Schema-Evolutions-Prozess werden auch Multitype-Schema-Operationen identifiziert, dessen Schemas verschieden sind. Der vorgestellte datengetriebene Schema-Evolutions-Prozess besteht aus den Verfahren Domänenklassifikation, Ähnlichkeits- und Verbundberechnung.

Eine Domäne eines Attributs ist eine Abstraktion der Attributswerte in vordefinierten Klassen und ist mit dem Datentyp eines Attributs vergleichbar. Im Gegensatz zum Datentyp ist die Domäne nicht abhängig von einer konkreten Implementierung durch einen Datenbankhersteller, sondern ist an die mathematischen Skalenniveaus gekoppelt [4][vgl. 2]. Für NoSQL-Datenbanken ist ein Schema durch die Domänenklassen der Attribute einer Entität bestimmt. Dies ermöglicht die Integrierung beliebiger NoSQL-Datenbanken in den Schema-Evolutions-Prozess, da das Schema kein datenbankabhängiges Konstrukt ist. Mithilfe eines Ähnlichkeitsmaßes werden Attribute durch das Teilen einer gemeinsamen Semantik zusammengeführt. Die Idee ist, dass eine neue Ähnlichkeitsbeziehung in den Daten erzeugt wird, wenn eine Multitype-Schema-Operation durchgeführt wurde. Wird für ein Attribut eine Ähnlichkeitsbeziehung nachgewiesen, muss ein Entitätsverbund berechnet werden. Ein Entitätsverbund beschreibt eine semantische Beziehung zwischen zwei Entitäten. Abschließend können die durchgeführten Multitype-Schema-Operationen isoliert werden.

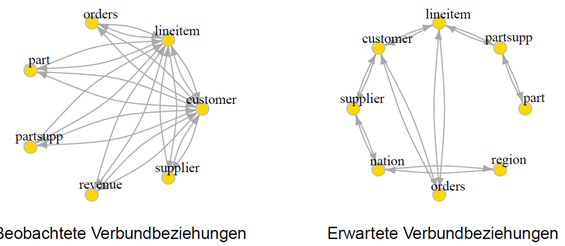
Ergebnisse

Für die Evaluation der Korrektheit der Ähnlichkeits und Verbundberechnungen werden die Ähnlichkeits- und Verbundbeziehungen in einem Graphen visualisiert. Ein Graph wird durch eine Knoten- und Kantenmenge definiert. Die Knotenmenge umfasst alle Entitäten eines Datenmodells. Eine Verbindung zwischen zwei Knoten wird hinzugefügt, wenn zwischen zwei Attributen unterschiedlicher Entitäten eine Ähnlichkeits- oder Verbundbeziehung besteht. So kann übersichtlich nachvollzogen werden, ob zwischen Entitäten Ähnlichkeiten oder Verbünde fehlen:



Beobachtete Ähnlichkeitsbeziehungen — Erwartete Ähnlichkeitsbeziehungen

Figure: Ähnlichkeitsgraph des TCP-H-Datenmodells



Beobachtete Verbundbeziehungen — Erwartete Verbundbeziehungen

Figure: Verbundgraph des TCP-H-Datenmodells

Die beobachteten Ähnlichkeits- und Verbundbeziehungen decken sich mit dem erwarteten Datenmodell. Abweichungen können durch fehlende oder fehlerhafte Daten erklärt werden. Anhand des Graphen kann festgestellt werden, dass wesentliche Kanten vorhanden sind. Dies spricht für eine gute Ermittlung von Ähnlichkeiten und Verbünde.

Die Ermittlung von Multitype-Schema-Operationen ist ein allumfassender Test der Verfahren. Es werden Multitype-Schema-Operationen durchgeführt (D). Diese werden anschließend aus den Daten extrahiert (E).

Commerzbank-Datenmodell:

(D) MOVE exp_gp.gp_status to exp_gp_organkredit.status WHERE exp_gp.gpkenn = exp_gp_organkredit.gpkenn

(E) Move exp_gp_v1.gp_status TO exp_gp_organkredit_v2.satus WHERE lhs.gpkenn = rhs.gpkenn

TPC-H-Datenmodell:

(D) MOVE exp_gp.gp_status to exp_gp_organkredit.status WHERE exp_gp.gpkenn = exp_gp_organkredit.gpkenn

(E) Move exp_gp_v1.gp_status TO exp_gp_organkredit_v2.gp_status WHERE lhs.gpkenn = rhs.gpkenn

Die erzeugten Multitype-Schema-Operationen konnten in beiden Datenmodellen zuverlässig erkannt werden.

Fazit

Das Ziel, die Multitype-Schema-Operationen *Move* und *Copy* datengetrieben zu erkennen, konnte nachweislich erfüllt werden. Die Verfahren sind nicht proprietär auf eine konkrete Implementierung eines Datenbankherstellers zugeschnitten. Es besteht die Möglichkeit, mit geringem Aufwand ein Datenbanksystem gegen ein anderes auszutauschen. Das Thema Schema-Evolution ist ein bedeutendes Thema für die Commerzbank. Alle Schema-Veränderungen müssen protokolliert und nachvollziehbar gemacht werden. Datenkonsumenten erhalten so eine stabile Grundlage für eine erfolgreiche Zusammenarbeit.

Danksagung

Ich bedanke mich herzlichst bei der Commerzbank AG und dem BDAA-Team für ihre tolle Unterstützung.

Literatur

[1] TPC,Transaction Processing Performance Council TPC (2011),1-134

[2] Thorsten Papenbrock und Felix Naumann. A Hybrid Approach to Functio-nal Dependency Discovery (2002)

[3] Andreas Meier und Michael Kaufmann. NoSQL Databases. SQL and NoS-QL Databases (2019)

[4] Chua Cecil Eng H. and Chiang and Roger H. L. and Lim Ee-Peng, Instance-based attribute identification in database integration (2003)

[5] Daniel Müller. Schema-Extraktion zur Unterstützung des Schema-Managements in NoSQL-Datenbanksystemen. Hochschule Darmstadt (2016)

[6] Meike Klettke, Uta Störl, Manuel Shenavai und Stefanie Scherzinger. NoSQL schema evolution and big data migration at scale. IEEE International Conference on Big Data. (2016)