Bachelorprojekt

Wintersemester 2015/2016 – Sommersemester 2016

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Verteilte Graphdatenbanken in der Cloud

Projekthintergrund

Graphdatenbanken gewinnen, z.B. für Anwendungen im Bereich von sozialen und wissenschaftlichen Netzwerken, Mediendatenbanken und Inferenzsystemen, an zunehmender Popularität. Im Gegensatz zu relationalen Datenbanken unterliegen die Inhalte von graphbasierten Datenbanken einer Graphstruktur und benutzen spezialisierte Algorithmen, um Suchanfragen effizient zu bearbeiten. Die Suchanfragen können über Grapherweiterungen für SQL modelliert werden und müssen durch geeignete Graphalgorithmen implementiert werden, um eine hohe Performanz zu erreichen.

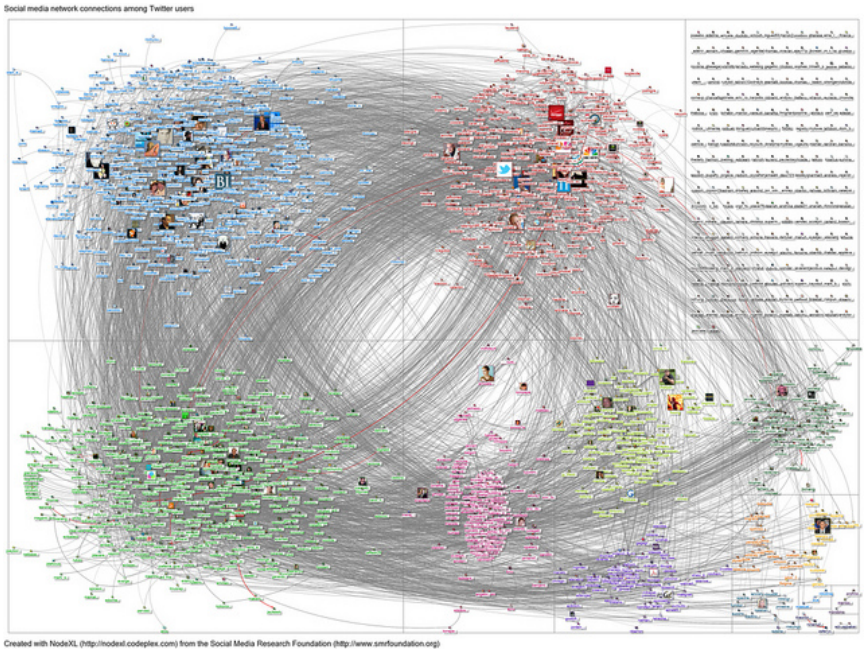

Abbildung 1 Ausschnitt aus dem Twitter Netzwerk für einen Nutzer[1]

Während Graphen in der Größenordnung von bis zu 10^7-10^8 noch effizient im Hauptspeicher einer modernen Serverarchitektur bearbeitet werden können, müssen für Graphen in der Größenordnung von Milliarden Knoten (wie z.B. das gesamte Twitter Netzwerk, Ausschnitt siehe Abb. 1) verteilte Datenbanksysteme eingesetzt werden. Neben der Ausnutzung von In-Memory Technologien und Multi-Core Architekturen in den einzelnen Knoten der Datenbank ist dabei die Entwicklung von Cloud-basierten verteilten Graphdatenbanken, die geeignete verteilte Datenstrukturen und Algorithmen nutzen, essentiell. Diese drei Richtungen in einer vereinheitlichten Architektur effizient zu integrieren, stellt dabei eine große Herausforderung dar.

In diesem Bachelorprojekt sollen zusammen mit dem SAP Innovation Center basierend auf den Theorien von Graphalgorithmen/Graphtransformationen und verteilten Systemen Erweiterungen einer in der Entwicklung befindlichen Graphdatenbank konzipiert und prototypisch implementiert werden. Dabei wird auf eine sich bei SAP derzeit in der Entwicklung befindende Graphdatenbank-Software aufgebaut. Für die Erweiterungen spielen existierende fortschrittliche Algorithmen für Graph-Pattern-Matching und Graph-Clustering eine wichtige Rolle. Desweiteren ist ein systematischer Ansatz zum Testen und Benchmarken der entwickelten Erweiterungen der verteilten Graphdatenbank-Software zu entwickeln.

---

[1] http://theincidentaleconomist.com/wordpress/wp-content/uploads/2014/06/NodeXL-Twitter-User-jowyang-network-graph.jpg

Projektgegenstand

Ziel dieses Bachelorprojektes ist die Konzeption und Implementierung von mehreren Erweiterungen einer graphbasierten verteilten Datenbank. Als erste Erweiterung sollen existierende fortgeschrittene Algorithmen für Graph-Pattern-Matching analysiert und prototypisch implementiert werden. Dazu ist ein Vergleich mit der Funktionalität in existierenden Graphdatenbanken, wie z.B. Neo4j[2], nötig. Als zweite Erweiterung sollen auch fortgeschrittene existierende Algorithmen für Graph-Clustering miteinander und mit anderen Methoden verglichen und schließlich prototypisch implementiert und evaluiert werden. Darüber hinaus soll ein Test- und Benchmark-Framework für verteilte Graphdatenbanken konzipiert und implementiert werden.

Umsetzung

Die ersten Schritte im Bachelorprojekt umfassen die Einarbeitung in die Themen Graphdatenbanken und Graphalgorithmen/Graphtransformationen sowie relevante Dokumentation. Desweiteren wird eine Einführung in die bei SAP befindliche Datenbankarchitektur gegeben und sowohl formale als auch technische Grundlagen für Erweiterungen gegeben. Nach einer detaillierten Anforderungsanalyse werden für die zu entwickelnden Teile passende Designs entwickelt und prototypisch implementiert. Abschließend wird das Konzept mit Hilfe bereitgestellter Testdaten evaluiert.

Projektumfeld

Das Projekt wird in Zusammenarbeit mit dem SAP Innovation Center in Potsdam stattfinden. Dort befindet sich derzeit eine Graphdatenbank in Entwicklung welche als Grundlage für die zu entwickelnden Erweiterungen dient.

Organisation

In der Seminarphase werden durch die teilnehmenden Studenten die Grundlagen zur genutzten Graphdatenbank-Software und zu ausgewählten Themen erarbeitet und präsentiert (1. Meilenstein). Die Ergebnisse der Seminarphase bilden die Grundlage für die folgende Anforderungserhebung, die in einem Anforderungsdokument zusammengetragen wird (2. Meilenstein). Das Anforderungsdokument wird in einem Antrittsvortrag dem SAP Innovation Center vorgestellt. Auf Basis der Anforderungen werden dann entsprechende Konzepte erarbeitet, die in einem Designdokument beschrieben werden (3. Meilenstein). Die Umsetzung der Konzepte wird in Form der Bachelorarbeiten beschrieben und evaluiert (4. Meilenstein). Abschließend werden die Ergebnisse des Bachelorprojektes in Form eines Abschlussvortrags vor dem SAP Innovation Center präsentiert (5. Meilenstein). Bei gutem Gelingen wird eine Anstellung als studentische Hilfskraft in Aussicht gestellt.

Teilnehmer und Projektbeginn

Bis zu 6 Teilnehmer können in diesem Projekt mitarbeiten. Projektbeginn ist der 1.10.2015.

Informationen

Für ausführliche Informationen zu dem Projekt stehen Prof. Holger Giese (A-2.5, holger.giese@hpi.de), Senior Researcher Dr. Leen Lambers (leen.lambers@hpi.de), Thomas Beyhl (A-2.11, thomas.beyhl@hpi.de), Johannes Dyck (A-2.8, johannes.dyck@ hpi.de), und Sebastian Wätzoldt (sebastian.waetzoldt@hpi.de).de und zur Verfügung. Ansprechpartner seitens des SAP Innovation Centers ist Dr. Christian Krause (SAP AG, christian.krause01@sap.com).

SAP Innovation Center

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

[2] Neo4j Homepage: http://neo4j.com/