Big-Data-Anwendungsentwicklung mit SQL und NoSQL

Jens Albrecht
Technische Hochschule Nürnberg
jens.albrecht@th-nuernberg.de

Uta Störl
Hochschule Darmstadt
uta.stoerl@h-da.de

Überblick

Bei der Verarbeitung und Auswertung von Big Data stoßen klassische relationale Datenbanken an ihre Grenzen. So ist die Speicherung von Daten im Multi-Terabyte-Bereich damit zwar möglich, aber in der Regel aufgrund der Lizenz- und Speicherkosten unwirtschaftlich. Datenanalysen, bei denen große Datenbereiche gelesen werden müssen, erfordern darüber hinaus cluster-fähige Systeme für die Skalierung.

Neben dem Datenvolumen stellen die Vielfalt der Daten und die Dynamik von Schemaänderungen ein Problem für relationale Datenbanken dar. Insbesondere Web-Anwendungen sind, neben hohem Datendurchsatz, durch sehr kurze Release-Zyklen geprägt. Schemaänderungen müssen im laufenden Betrieb durchgeführt werden, da Downtimes meist inakzeptabel sind.

Um die genannten Schwachstellen zu adressieren, haben sich in den letzten Jahren einige neue Datenbank-Ansätze etabliert. Zu nennen sind hier zum einen NoSQL-Datenbanken, die sich in der Regel durch ein sehr einfaches, aber flexibles Datenmodell und sehr gute Skalierbarkeit auszeichnen. Zum anderen hat sich für die Speicherung großer Datenmengen sowie deren Aufbereitung und Analyse das Apache-Framework Hadoop als Standard etabliert.

Das Tutorium legt den Schwerpunkt auf zwei relativ neue Aspekte dieser Technologien, die jeweils in einem 90-minütigen Vortrag vorgestellt werden:

1. Datenanalyse mit SQL auf Hadoop
2. NoSQL-Anwendungsentwicklung mit Objekt-NoSQL Mappern

Im Folgenden werden diese Themen kurz vorgestellt.

Datenanalyse mit SQL auf Hadoop

Hadoop ist ein Framework zur Speicherung und Verarbeitung großer Datenmengen in einem Cluster aus Standard-Servern. Es besteht im Kern aus dem verteilten Hadoop-

Filesystem sowie dem Map-Reduce-Framework. Letzteres stellt eine Ablaufumgebung für die verteilte Ausführung von Jobs, die in Java programmiert werden müssen, dar.

Die Programmierung von Map-Reduce-Jobs für die Datenanalyse ist, verglichen mit SQL, sehr aufwändig und das Know-How dafür in den wenigsten Firmen vorhanden. Andererseits sind auch bei Big Data viele Abfragen inhaltlich wenig komplex, wie folgende Beispiele illustrieren:

- Ermittlung der Häufigkeiten der Seitenaufrufe (Visits) für bestimmte Web-Sites
- Ermittlung durchschnittlicher Werte von Sensordaten pro Messstation
- Suche nach bestimmten Hash-Tags in Twitter-Nachrichten

Diese und ähnliche Abfragen lassen sich sehr einfach mit SQL ausdrücken. Notwendig dafür ist ein relationales Datenmodell, das für viele Daten, die satzweise erzeugt werden und eine gewisse Struktur aufweisen, einfach definiert werden kann. Eine verteilte SQL Execution Engine sorgt dann für die Ausführung im Hadoop-Cluster.

Bereits im Jahr 2007 wurde zu diesem Zweck Hive entwickelt, eine SQL-Engine, die Map-Reduce-Jobs auf Hadoop generiert. In den vergangenen zwei Jahren ist eine Vielzahl weiterer SQL-Systeme für Hadoop auf den Markt gekommen. Anders als Hive setzen Produkte wie Cloudera Impala oder Pivotal Hawk auf eigene Execution Engines anstelle Map-Reduce, um eine deutlich bessere Performance zu erreichen. Parallel dazu arbeiten die großen Datenbank-Hersteller Oracle, IBM und Microsoft an Lösungen, um Hadoop bestmöglich in ihre Produkte zu integrieren. Ganz neue Möglichkeiten eröffnet Apache Spark, ein In-Memory-Framework zur Datenverarbeitung. Mit Apache Spark SQL gibt es auch dafür bereits einen SQL-Aufsatz.

Der Vortrag gibt einen Überblick über existierende SQL-Zugriffsmöglichkeiten auf Daten in Hadoop und stellt die verschiedenen Architekturansätze anhand einer Klassifikation gegenüber.

NoSQL-Anwendungsentwicklung mit Objekt-NoSQL Mappern

Bei der Anwendungsentwicklung für NoSQL-Datenbanken gibt es verschiedene Herausforderungen: Da sich die Anfragesprachen für NoSQL-Systeme erheblich unterscheiden, führt eine native Anwendungsentwicklung unmittelbar zu starken Abhängigkeiten von einem spezifischen System. Außerdem führt die Schemaflexibilität der NoSQL-Datenbanken dazu, dass die Verantwortung für das Schema-Management in die Anwendung verlagert wird.

Einen möglichen Lösungsansatz für diese Herausforderungen stellen Objekt-NoSQL Mapper dar. Diese Systeme übertragen die Idee der Objekt-Relationalen Mapper auf NoSQL-Systeme. Aktuell finden sich am Markt dabei sowohl Produkte aus dem Bereich der Objekt-Relationalen Mapper, welche um die Unterstützung für (einige) NoSQL-Systeme erweitert wurden (beispielsweise EclipseLink, Hibernate und DataNucleus) als auch speziell

für den Bereich des Objekt-NoSQL Mapping entwickelte Systeme wie Kundera. Sofern es sich um Objekt-NoSQL Mapper für Java handelt, setzen fast alle diese Produkte auf die Java Persistence Language (JPQL) aus der Java Persistence API (JPA) als einheitliche Anfragesprache.

Aufgrund der größeren Flexibilität der Abbildung eines Objektmodells auf NoSQL-Datenmodelle und der sehr unterschiedlichen Mächtigkeit der NoSQL-Anfragesprachen unterscheiden sich die existierenden Produkte derzeit allerdings sowohl konzeptionell als auch bezüglich der Unterstützung verschiedener NoSQL-Systeme erheblich.

Im Tutorium wird ein Überblick über die verschiedenen existierenden Ansätze gegeben, diese werden kategorisiert und miteinander verglichen. Eine weitere wichtige Frage beim Einsatz von Mappern ist die Auswirkung auf die Performance der Lese- und Schreibzugriffe. Auch hierfür werden im Tutorium Untersuchungen präsentiert und Einsatzempfehlungen für Objekt-NoSQL Mapper gegeben.

Zielgruppe

Zielgruppe des Tutoriums sind Praktiker und Anwender, die sich über verschiedene Architekturkonzepte und den aktuellen Stand in Produkten zur Big-Data-Anwendungsentwicklung informieren wollen.

Die Vortragenden

Jens Albrecht ist seit 2014 Professor für Datenbanken und Big Data an der Technischen Hochschule Nürnberg, nachdem er bereits zwei Jahre an der Hochschule Würzburg-Schweinfurt eine Professur inne hatte. Zuvor war er als Berater bei Oracle und als WH-Architekt in der GfK in Nürnberg tätig. Sein besonderes Interesse gilt Technologien zur Aufbereitung und Analyse großer Datenmengen.

Uta Störl ist seit 2005 Professorin für Datenbanken und Informationssysteme an der Hochschule Darmstadt. Aktuelle Schwerpunkte ihrer Arbeit sind Technologien für NoSQL-Datenbanksysteme sowie interdisziplinäre Projekte im Bereich Digital Humanities. Nach ihrer Promotion 1999 an der Friedrich-Schiller-Universität in Jena hat sie im Research- bzw. Projektmanagement-Bereich der Dresdner Bank in Frankfurt am Main gearbeitet.