

Anfragesprachen für BigData

BigData ist ein Schlagwort für sehr große, sehr komplexe, sehr schnelllebige oder unstrukturierte Daten. Der Begriff umfasst auch intime und private Daten. Beispielsweise haben Smart Speaker wie Amazon Alexa eine ständige Verbindung zum Internet. Damit diese auf Spracheingaben reagieren können, muss das Mikrophon dauerhaft eingeschaltet sein und überträgt alle damit aufgenommenen Geräusche in die Cloud des Herstellers, wo die Geräusche auf darin enthaltene, an das Gerät gerichtete Kommandos, ausgewertet werden. Hierbei werden dementsprechend auch vertrauliche Gespräche mit in die Cloud übertragen und der Nutzer kann sich nie sicher sein, dass diese nicht auch aufgezeichnet werden.

Um die Auswertung der riesigen Datenmengen, die in einer solchen Cloud gespeichert werden, zu vereinfachen, existieren verschiedene Anfragesprachen. Einige davon, auf die ich näher eingehen werde, sind Cypher, die Drools Rule Language und SQL++.

Cypher ist eine Abfragesprache, die speziell für die Graphdatenbank Neo4j entwickelt wurde. Je nach Anwendungsfall kann Neo4j für BigData eingesetzt werden, wenn es sich um Daten handelt, die stark miteinander verknüpft sind bzw. keine feste Struktur besitzen. Beispielsweise trifft dies auf Straßennetze und Beziehungsnetzwerke zu. Im Bereich des autonomen Fahrens sind sehr genaue Straßenkarten notwendig, um eine zuverlässige Navigation der Fahrzeuge zu ermöglichen.

Drools Rule Language ist die Abfrage des Geschäftsregel-Management-Werkzeugs Drools. Im Rahmen von BigData ist der Einsatz von Drools sinnvoll, wenn ein großer Datensatz aus vielen Teildaten vorhanden ist, für die keine komplexen Zusammenhänge bekannt sind, die sich aber durch kleine, elementare Regeln beschreiben lassen. Beispielsweise ließe sich mit Drools prinzipiell eine Wettersimulation realisieren. Es kann keine komplexe Regel aufgestellt werden, die beschreibt, welchen Einfluss das Klima in Amerika auf das Klima in Europa hat. Bekannt sind aber die lokalen Wetterbedingungen an den Messpunkten und die Art und Weise der Wechselwirkung mit den benachbarten Messpunkten. Wenn sich die Temperatur an einem Messpunkt erhöht, dann wird sich in abgeschwächter Form auch die Temperatur an den benachbarten Messpunkten erhöhen. Und wenn nun die elementaren Regeln auf alle Messpunkte angewendet werden, dann kann man so simulieren, welches Wetter am nächsten Tag herrschen wird.

Bei SQL++ handelt es sich um reguläres SQL, das dahingehend erweitert wurde, dass nun JSON-Objekte als Resultate zurückgegeben werden. Damit lassen sich in einer gewöhnlichen Datenbank nun auch ganze Objekte abspeichern. Der Einsatz von SQL++ ist im BigData-Umfeld dann sinnvoll, wenn mittels REST-API der Zugriff auf die Daten ermöglicht werden soll, da hier Format-Transformationen zwischen Ausgabe der Datenbank und REST-Schnittstelle vermieden werden können.

Links

<https://www.heise.de/newsticker/meldung/Googles-Sprachassistent-Mitarbeiter-hoeren-und-bewerten-Audioaufnahmen-4467985.html>

<https://www.elektronikpraxis.vogel.de/mit-hochpraezisen-hd-karten-wird-autonomes-fahren-machbar-a-652537/>