

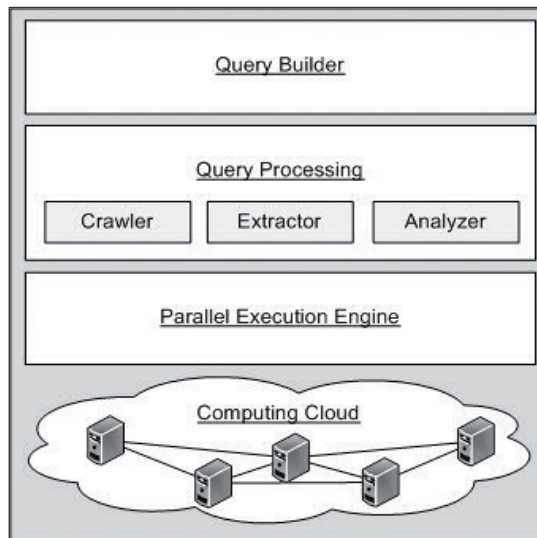
Interview mit Prof. Dr. Markl, TU Berlin: Neue Forschung im Bereich Business Intelligence

Informationsmanagement im Umbruch

„Challenges in Information Management“ lautet der Vortrag von Prof. Dr. Markl, der mit diesem Thema die datacon '09 [1] im November eröffnen wird. Im Vorfeld sprachen wir mit Prof. Dr. Markl, der das Datenbank-Business auch von Seiten der Industrie kennt (siehe Kasten), über die momentan größten Herausforderungen in Lehre und Forschung. Das Gespräch für database pro führte Prof. Stefan Edlich, der zum Advisory Board der datacon '09 gehört und auf der Konferenz im November den Track „Cloud Databases“ moderieren wird.

database pro: *Wo sehen Sie die großen Herausforderungen im Informationsmanagement der Zukunft?*

Markl: Informationsmanagement wird von zwei Seiten mit neuen Herausforderungen konfrontiert: Zum einen gibt es regelmäßig neue Herausforderungen, die durch neue Hardware- und Software-Infrastrukturen entstehen. Zum anderen ergeben sich durch neuartige Anforderungen seitens der Anwender große Herausforderungen. Wir erleben derzeit mehrere Umwälzungen im Bereich der Hardware- und Software-Infrastrukturen sowie im Bereich der Geschäftsmodelle. Stichworte in diesen Bereichen sind insbesondere Multi-Core, Text Mining und Cloud Computing. Lassen Sie mich diese exemplarischen Themen etwas erläutern: Beispielsweise haben wir uns im vergangenen Jahrzehnt an eine regelmäßige Verdoppelung der Taktfrequenz von Mikroprozessoren in jeder neuen Prozessorgeneration gewöhnt – und damit an eine ständige Erhöhung der Anzahl an Berechnungen, die



Architektur eines Situational-Business-Intelligence-Systems

ein Prozessor in einer Sekunde durchführen kann. Dieses exponentielle Wachstum der Verarbeitungsgeschwindigkeit, welches oftmals auch als „Moore’s Gesetz“ bezeichnet wird, hat die rasante Entwicklung von komplexen Softwaresystemen – und damit Informationsmanagementanwendungen – überhaupt erst ermöglicht. Allerdings ist dieses „Gesetz“ in den letzten Jahren aufgrund der erforderlichen ständigen Verkleinerung der Chipstrukturen an technische Grenzen gestoßen. Prozessorhersteller wie Intel, AMD oder IBM bauen in den letzten Jahren daher mehrere, parallele Rechenwerke in einen Chip ein, um die Leistung der Prozessoren auch weiterhin zu steigern. Leider unterstützt heutige Software derartige „Multi-Core“ Prozessoren nur unzulänglich. Somit werden in den nächsten Jahren im Bereich der Informationsmanagementsysteme viele Forschungs- und Entwicklungsarbeiten im Bereich der Parallelisierung von Anwendungen auf Multi-Core-Strukturen stattfinden. Ebenso werden neue Speicherformen wie SSD eine nachhaltige Änderung der Datenbanksystemarchi-

tektur verursachen. Datenbanksysteme optimieren im Wesentlichen den Flaschenhals zwischen Speicher und Festplatte. Allerdings verändern sich die Eigenschaften dieses Flaschenhals beim Einsatz von SSD anstelle von Festplatten: SSD erfordern keine Positionierung eines Schreib/Lese-Kopfes über einer Magnetplatte und haben daher ein völlig anderes Verhalten beim wahlfreien Zugriff auf Daten, die an einer beliebigen Stelle auf einer SSD abgespeichert sind. Neben diesen Änderungen der Hardware- und Software-Infrastruktur ergeben sich neue Anforder-

ungen durch die Etablierung eines neuen Geschäftsmodells, welches Software nicht mehr als Produkt verkauft, sondern als Dienstleistung anbietet. Dieser Trend, der oftmals auch als „Cloud Computing“ bezeichnet wird, wirft neue Fragestellungen auf, zum einem im Bereich der Software-Infrastruktur, zum anderen aber auch in sozio-ökonomischen und juristischen Bereichen, wie Vertrags-

recht, Haftung, Datenschutz und Datensicherheit. Allerdings sind auch die Herausforderungen durch neuartige Anwendungen beträchtlich. Aufgrund der Vielfalt von Textdaten im Web, in Unternehmensnetzen, in E-Mails, Wikis, Blogs und Foren zeichnet sich als ein neuer Trend die Analyse von Texten, sogenannten unstrukturierten Daten, ab. Diese Textdaten müssen zusätzlich zu klassischen strukturierten Daten, wie sie in relationalen Datenbanksystemen gespeichert werden, analysiert werden. Beispiel für eine derartige Textanalyse wäre es, herauszufinden, wie viele Nutzer sich in einem Webforum positiv über ein bestimmtes Produkt äußern. In vielen Anwendungen sollen sogar die Textdaten mit den klassischen strukturierten Daten verknüpft werden. Beispiel hierfür wäre es, die Effektivität einer Marketingkampagne mit den Äußerungen über dieses Produkt in einem sozialen Netzwerk zu korrelieren. Meine Prognose

SSD ändert die Datenbankarchitektur der Zukunft

Prof. Dr. rer. nat. Volker Markl leitet das Fachgebiet Datenbanksysteme und Informationsmanagement (DIMA) an der Technischen Universität Berlin. Vor seiner Tätigkeit in Berlin leitete er Forschergruppen am Bayerischen Forschungszentrum für wissensbasierte Systeme (FORWISS) sowie am IBM Almaden Forschungszentrum in San Jose, CA, USA. Volker Markl promovierte an der Technischen Universität München. Seine Forschungsinteressen beinhalten Dienste und neue Rechnerarchitekturen für das Informationsmanagement sowie Indexierung, Anfrageverarbeitung, Optimierung Informationsextraktion und Informationsintegration.



AKTUELL

ist, dass die nächste Generation von Business-Intelligence-Anwendungen neue Funktionen bereitstellen wird, welche die automatische Analyse großer Mengen von Textdaten ermöglichen. Daneben wird die nächste Generation von BI-Anwendungen nicht nur existierende Daten analysieren, sondern auch Funktionen zur Vorhersage und Planung beinhalten.

database pro: Können Sie Beispiele für typische Anfragen im Bereich Business Intelligence geben, die heute schwer zu beantworten sind?

Markl: BI-Anwendungen können heutzutage sehr gut die Daten innerhalb eines Unternehmens verarbeiten, insofern diese durch einen ETL-Prozess in ein Data Warehouse überführt wurden. Erheblich schwieriger ist die Korrelation dieser unternehmensinternen Daten mit externen Datenquellen, zum Beispiel eine Gegenüberstellung der Verkäufe in einer Region mit den Verkäufen von Konkurrenzprodukten. Noch schwieriger wird es, wenn, wie oben beschrieben, die Datenquellen nicht strukturiert sind, sondern eine Menge von Texten sind. Typisches Beispiel dafür ist es, herauszufinden, wie die Stimmung für mein Produkt derzeit bei den Kunden ist. Man könnte ein derartiges „Sentiment“ durch die Analyse von Testberichten, Nachrichtenartikeln oder sogar Blogs im Internet herausfinden. Auf diese Weise können Unternehmen dann Markenpflege betreiben. Ähnliche Analysen sind für die Planung einer Produkteinführung und zur Bewertung von Marketingkampagnen sinnvoll. Allerdings erfordern derartige Analysen neuartige Verfahren zur Informationsextraktion aus Texten, eine hochgradig effiziente Infrastruktur zur Verarbeitung von großen Datenmengen und Techniken, um mit der Unsicherheit umzugehen, die beispielsweise bei der Informationsextraktion auftritt.

Business Intelligence

Unter Business Intelligence versteht man Verfahren und Methoden zur systematischen Analyse von Daten. Dies sind meist Unternehmens- oder Marktdaten, die so ausgewertet werden sollen, dass bessere strategische Entscheidungen getroffen werden können. In der Regel sind dies immer bessere und komplexere Reports, die oft auch aus einem Data Warehouse gewonnen werden können.

database pro: Welche Arbeiten verfolgen Sie im Rahmen von neuen Business-Intelligence-Lösungen?

Markl: Meine Gruppe an der Technischen Universität Berlin untersucht derzeit, wie Business Intelligence unter Einbeziehung von Textdaten effizient auf einer massiv parallelen Rechnerinfrastruktur ausgeführt werden kann. Dadurch wollen wir grundlegende Technologien und Infrastruktur entwickeln, um die oben beschriebenen Analysen zu ermöglichen. Wir betrachten dabei neuartige Verfahren, um Daten auf einer adaptiven, verteilten Rechnerinfrastruktur (also einem Cloud-System) zu speichern, zu aggregieren, zu kombinieren und zu analysieren. Ziel ist es dabei, Anfragen über großen Textdatensmengen, wie zum Beispiel einem Teil des Internet, verarbeiten zu können. Beispielsweise haben wir ein System prototypisch implementiert, welches die Frage „Was ist das Durchschnittsalter von Geschäftsführern von Unternehmen je nach Land“ durch Analyse einer Menge von Nachrichtentexten automatisch ableiten kann. Klassische Suchmaschinen können eine derartige Anfrage nur beantworten, falls diese Statistik manuell von einem Benutzer erstellt wurde und auf eine Webseite ins Internet gestellt wurde. Unser Prototyp hingegen setzt dies nicht voraus, sondern kann die Information durch Kombination und Aggregation von vielen Dokumenten ableiten. Allerdings sind, abhängig von Datenquelle und Informationsextraktionsverfahren, die Unsicherheiten, die im Rahmen der Anfrageverarbeitung auftreten – und damit die Qualität der Antwort – noch eine Herausforderung. Daher muss man die Datenbasis, die man für die Analyse benutzt, verstehen beziehungsweise versuchen, das Rauschen, welches durch fehlerhafte Daten entsteht, durch statistische Methoden zu verringern oder auszuschalten. Neben dem Thema „Business Intelligence über Textdaten“ arbeite ich zusammen mit Kollegen im Raum Berlin an einem Informationsmanagementsystem, welches derartige Anfragen effizient auf einer Cloud-Infrastruktur verarbeiten kann. Dabei setzen wir auf ein sogenanntes funktionales Programmiermodell (d.h. ein Modell, das auf Operationen aus funktionalen Programmiersprachen wie Map und Reduce basiert), welches die massive Parallelisierung von Anfragen auf Multicore-CPUs

beziehungsweise vielen Rechnern ermöglicht. Nur auf diese Weise können komplexe Anfragen in vernünftiger Zeit verarbeitet werden.

database pro: Reicht das „klassische“ Map/Reduce-Modell noch aus?

Markl: Map und Reduce sind zwei Operationen einer funktionalen Sprache, die sich sehr gut zur Parallelisierung von Filterung und Aggregation eignen, das heißt zur massiv parallelen Verarbeitung der SQL-Operationen „SELECT SUM() FROM single_table... GROUP BY WHERE ...“. Map/Reduce wurde durch Google und Yahoo populär sowie durch Hadoop, eine Open-Source-Implementierung eines Map/Reduce-Prozessors auf einem verteilten Dateisystem. Allerdings eignet sich klassisches Map/Reduce nicht wirklich zur Verknüpfung von mehreren Datenquellen, das heißt den Join der relationalen Algebra. Einige Open-Source-Projekte, wie zum Beispiel Apache Mahout, versuchen bereits, Data-Mining-Verfahren auf einem Map/Reduce-Modell auszu-drücken. Allerdings ergeben sich im klassischen Map/Reduce-Modell Schwierigkeiten bei Zeitreihenanalyse und komplexeren Maschinenlernalgorithmen beziehungsweise Data-Mining-Operationen.

database pro: Was ist für eine leistungsfähige Business-Intelligence-Lösung auf der Seite der Datenbank nötig, und welche Rolle spielt die Textanalyse dabei?

Markl: Die Verschmelzung von Textdatenanalyse mit strukturierter Datenanalyse erfordert neue Konzepte an vielen Stellen innerhalb der Architektur eines Business-Intelligence-Systems. Zum einen ist eine Anfragesprache nötig, die die Spezifikation von Analysen über Textdaten ermöglicht. Man kann sich vorstellen, dass die Dimensionen, die in einem klassischen Data Warehouse vordefiniert sind, sich in einem „BI über Text“-System durch Informationsextraktion erst ergeben. Dabei muss das System der Unsicherheit bei der Informationsextraktion Rechnung tragen, indem es die Qualität der Antworten, welche sich aus den verwendeten Quellen und Extraktionsverfahren ergibt, anzeigt. Ferner muss das System den Benutzer bei der Anfrageformulierung unterstützen, indem es mögliche weitergehende Anfragen vorschlägt. Dieser im Englischen „query refinement“ genannte Prozess führt zu einer explorativen Analyse, die gerade im Bereich der Textdatenanalyse wichtig ist.

[Stefan Edlich/ef]

[1] www.data-conference.de