In-Memory-Datenbanken: Quantensprung oder kurzlebiger Hype?

CHRISTOPH MEINEL, HASSO-PLATTNER-INSTITUT GMBH

Um riesige Datenberge sekundenschnell auszuwerten, bedienen sich Softwarehersteller zunehmend der sogenannten In-Memory-Datenbanken. Daten werden hierbei nicht auf der Festplatte, sondern im Hauptspeicher des Rechners vorgehalten und auch dort verarbeitet. Die Geschwindigkeitsvorteile ergeben sich durch den schnelleren Datenzugriff auf den Hauptspeicher. In-Memory-Datenbanken liefern schnelle und aussagekräftige Analysen, ohne dass die Daten hierfür besonders aufbereitet werden müssen, und sind damit besonders für Anwendungsbereiche wie Business Intelligence (BI) attraktiv. Mit dieser Technologie erhalten Anwender ihre Auswertungen tagesaktuell und im Prinzip auf Knopfdruck – das zumindest ist das Leistungsversprechen der Anbieter. Wir haben nachgefragt, bei Prof. Dr. Christoph Meinel, Institutsdirektor und CEO des Hasso-Plattner-Instituts (HPI).

Mutet der aktuelle Hype um In-Memory nicht etwas verwunderlich an, wo es diese Technologie grundsätzlich doch schon seit Jahrzehnten gibt? Warum schlägt das Thema gerade jetzt so hohe Wellen?

CM: Bei In-Memory geht es zunächst einmal darum, dass Daten nicht mehr auf langsamem Festplattenspeicher gehalten werden, sondern nur noch im Hauptspeicher selbst. Darüber hinaus werden die Daten so organisiert, dass sie sehr schnell ausgewertet werden können. Dem steht die Tabellenstruktur klassischer Datenbanken entgegen. Die Daten werden zeilenweise gespeichert. Wenn ich aber die Programme betrachte, die die Daten in den Tabellen auswerten sollen, dann ist es bei vielen Anwendungen so, dass sie in Spalten analysieren. Wenn die Tabellen dann zeilenweise abgespeichert sind, aber spaltenweise ausgewertet werden, muss man immer springen, um den nächsten Eintrag zu finden. Das heißt, dass dann die Rechenzeit sehr stark davon bestimmt ist, dass man den richtigen Eintrag findet. Bei In-Memory-Datenbanken werden die Daten im Speicher spaltenweise abgelegt, man spart also das Springen. In-Memory-Technologie ist dabei ein Konglomerat von Technologien, die es in Teilen schon gab, die aber erst jetzt bedeutsam werden, da die entsprechenden Hardwareentwicklungen gefolgt sind, mit denen man das umsetzen kann. Der Hauptspeicher ist der teuerste Speicher im Rechner, da er direkt am Prozessor lokalisiert ist. Früher war er, wegen des hohen Preises, sehr klein, die deutlich größere Festplatte ist viel preiswerter, aber viel weiter vom Prozessor entfernt. Der Transfer der Daten auf die Festplatte und zurück ist dabei 100.000 Mal langsamer als der Rechenprozess selbst. Bei der In-Memory-Technologie liegen, wie erwähnt, alle Daten in Hauptspeicher. Das Rechnen erfolgt jetzt wahnsinnig schnell, weil die Daten ja nun nicht mehr zwischen Festplatte und Hauptspeicher hin- und hergeschoben werden müssen. Zudem kommen Multicore-Core Prozessoren zum Einsatz, bei denen in einen Prozessor mehrere Prozessor-Kerne eingebaut sind. Das zusammen macht es möglich, um Größenordnungen schneller zu rechnen, und damit unvorstellbar riesige Datenmengen in Sekundenschnelle auszuwerten. Heute gibt es Rechner mit einem Hauptspeicher von 2 Terabyte – das ist eine Größenordnung in die alle Daten eines mittelständischen Unternehmens hineinpassen.

Fest steht, bei der Analyse von Massendaten kommt man künftig nicht an In-Memory-Technologie vorbei. Das macht ein Beispiel aus der Genom-Forschung deutlich, wo es wahrhaftig um Big Data geht. Nach dem biochemischen Teil einer Genom-Analyse, also der Sequenzierung der Materialschnipsel in ihrer Ausgangsbasenkonstruktion, kommt die IT ins Spiel: Dann müssen die Schnipsel richtig kombiniert werden, man nennt das Alignment. Hier sind aufwändige Rechenschritte gefordert, denn man muss eine riesige Zahl von Kombinationsmöglichkeiten analysieren. Mit Multicore Rechnern kann das in Zukunft sehr schnell erledigt werden. Wenn dann die Genomergebnisse vorliegen, und man etwa feststellt, dass an gewissen Stellen Mutationen aufgetreten sind, ergibt sich eine weitere Aufgabe. Nun muss in allen weltweiten Forschungsdatenbanken geprüft werden, ob diese Mutation einem Krankheitsbild entspricht oder mit einem Krankheitsbild verknüpft ist. Diese Daten sind weltweit verstreut und nicht in einem einheitlichen Datenformat verfügbar. Früher dauerte der geforderte Abgleich Wochen. Durch die Parallelisierung mit 1000 Core Rechner können die Daten aus allen Datenbanken in ein bis zwei Nächten in die In-Memory-Datenbank hereingeholt werden. Anschließend können die Daten in wenigen Sekunden ausgewertet werden. Das Verfahren ist grundsätzlich nicht neu, aber über die Möglichkeit, Daten im Hauptspeicher abzulegen und dort nach gewissen Mustern durchsuchen zu können, ergibt sich ein enormer

KURZ UND BÜNDIG

Kurz und bündig:
Bei In-Memory geht es grundsätzlich darum, dass Daten nicht mehr auf langsamem Festplattenspeicher gehalten werden, sondern nur noch im Hauptspeicher selbst. Darüber hinaus werden die Daten so organisiert, dass sie sehr schnell ausgewertet werden können Bei der Analyse von Massendaten wird man künftig nicht mehr an der In-Memory-Technologie vorbeikommen. Praxisbeispiele aus der Medizinforschung zeigen schon heute die vielfältigen Perspektiven und Möglichkeiten auf.

Geschwindigkeitsgewinn. Ich denke das Beispiel zeigt die Potenziale der In-Memory-Technik!

Noch stehen In-Memory-Datenbanken vielfach im Schatten klassischen Datenbanklösungen für operative Daten und Data Warehouses für analytische Daten. Wie wird die zukünftige Dateninfrastruktur von Unternehmen aussehen? Kommt es zur Verschmelzung von Datenbank- und Datawarehouse?

CM: In der Vergangenheit haben wir eine Trennung zwischen der Transaktion von Daten und ihrer Analyse gehabt. Die Transaktionen waren zwar schnell, aber man kam mit der Analyse der Daten nicht nach. Mittels In-Memory-Datenbanken lassen sich diese Aufgabentypen verschmelzen, sie können durch eine einzige Datenbank ausgeführt werden. Theoretisch muss es also in einem Unternehmen nur noch eine Datenbank geben. Wie schnell dieser Prozess vorankommt, wird sich zeigen. Man kann nicht von heute auf morgen alles umstellen, denn auch die Prozesse müssen den In-Memory-Datenbanken angepasst werden. Das braucht Zeit. Vermutlich wird es für eine gewisse Frist einen ganzen Zoo von Datenbanken geben. Für das Back-up von Daten und ihre langfristige Archivierung werden traditionelle Datenbanken sicher weiter ihre Bedeutung haben.

Abgesehen von der schnelleren Informationsverarbeitung, wie sehen konkrete Killerapplikationen aus, die der Technologie „In-Memory" zum Durchbruch verhelfen? Sind es nur analytische Anwendungen für klassische Unternehmensprozesse oder wo geht die Reise hin?

CM: In-Memory-Datenbanken eigenen sich nicht nur für große Analyseprojekte, wie etwa in der Gen-Forschung. Gerade die betrieblichen Analyseaufgaben können von der In-Memory-Technologie profitieren. Übrigens stand gerade dieser Einsatzbereich am Anfang der Entwicklung. Viele Leute können sich gar nicht den Umfang der betrieblichen Daten vorstellen, die es da zu analysieren gibt! Und wenn man sich vor Augen führt, dass in der Vergangenheit und in Teilen noch heute ein Management-Board zusammensitzt und Entscheidungen auf der Basis absolut veralteter Daten treffen muss, dann wird deutlich, wie wichtig der Zugriff auf Echtzeitdaten ist! Ich weiß von großen Firmen, dass sie aufgrund veralteter Datenlage falsche Entscheidungen getroffen haben. Ich kann aber auch von einer Fluggesellschaft berichten, die schon heute erfolgreich auf In-Memory-Technologie setzt. Dort ist ein Pricing im Sekundenbereich möglich. Je nach Marktlage, Verkaufssituation und Anzahl der noch verfügbaren Plätze wird in Sekundenschnelle ein angepasstes Preismodell entwickelt und umgesetzt – das war bisher undenkbar.

CM: Grundsätzlich geht es für moderne Unternehmen um die Verfügbarkeit aktueller Informationen auf Fingertipp. Das heißt, dass ich als Vorstand nicht mehr am Montag eine Frage stelle und am Freitag dann eine Antwort bekomme, sondern dass ich während einer Sitzung am Tablet Abfragen an meine Datenbanken stellen kann und in Realzeit, also sofort Antworten erhalte. Das macht einen ganz anderen Führungsstil möglich, ganz neue Entscheidungsprozesse werden möglich. Entscheidungen können jetzt sehr viel substanzieller getroffen werden.

Wie sieht es mit der Sicherheit der Daten aus, die nur im Arbeitsspeicher vorgehalten werden? Was passiert bei einem Systemabsturz?

CM: Wenn der Strom weg ist, sind die Daten im Hauptspeicher tatsächlich erst einmal weg. Das bedeutet, man muss sich ganz genau überlegen, wie man das verhindert, bzw. wie man sicherstellt, dass diese Daten weiter zur Verfügung stehen. Das ist jetzt die Kunst, wie man aus so einer Idee der In-Memory-Datenverarbeitung ein leistungsfähiges In-Memory-Datenbankprodukt baut. Eine elementare Anforderung besteht darin, dass die Daten in kurzer Zeit wieder zur Verfügung stehen, selbst wenn etwas passiert. Die HANA Datenbank der SAP wäre nicht verkaufbar, wenn man keine Lösungen gefunden hätte, um Daten zu rekonstruieren. Da kommt auch das Thema Archivierung wieder ins Spiel. Grundsätzlich gilt, Hauptspeicherdatenbanken sind bei entsprechender Konfiguration im Ergebnis nicht unsicherer, als klassische Datenbanksysteme.

Auch mit Blick auf die Datensicherheit im Sinne der Vertraulichkeit kann ich aus der Erfahrung des HPI beruhigen. Das HPI verarbeitet in seinem Future SOC Lab mit In-Memory-Datenbanken Daten für verschiedene Forschungspartner und das funktioniert einwandfrei. Über Betriebssysteme ist sichergestellt, dass sich die Daten verschiedener Nutzer

weder vermischen noch abfließen. Das ist übrigens auch bei traditionellen Datenbanken sichergestellt. Sicherheit ist meist kein technisches Problem, sondern ein Problem der handelnden Menschen. Die Prozesse selbst können nicht durcheinanderkommen. Es sind Menschen, die Daten entwenden oder missbrauchen.

Anbieter werben damit, dass große Datenmengen bis zu 3600 Mal schneller analysiert, parallele Verarbeitungsvorgänge und Berechnungen schnell ausgeführt werden können und auch Kosten durch eine vereinfachte IT-Landschaft gesenkt werden. Sind hier gute Marketingstrategen unterwegs, oder öffnen sich in der Tat ganz neue Zukunftsperspektiven?

CM: Ich denke die Geschwindigkeitspotentiale bei der Datenanalyse stehen außer Zweifel. Eine Vereinfachung entsteht in dem Sinne, dass nicht mehr zwei unterschiedliche Datenbanken gebraucht werde um Transaktionen und Analysen durchzuführen. Die Frage ist natürlich, wie schnell solche Vereinfachungen greifen werden. Das gilt für Unternehmen gleichermaßen wie für Rechenzentren. Ich kann für die Forschungslandschaft bestätigen, dass die neue Technologie dank ihrer Möglichkeiten breit angenommen wird.

In der interdisziplinären Forschungsinitiative HANA Oncolyzer haben sich das Institut für Pathologie der Charité-Universitätsmedizin Berlin, das SAP Innovation Center Potsdam und das Hasso-Plattner-Institut dem Ziel verschrieben, die bislang vorrangig für Geschäftssoftware verwendete In-Memory-Technologie auch für den medizinischen Einsatz nutzbar zu machen. Welche Ziele verfolgen Sie dabei und welche weiteren Anwendungsszenarien könnten sich aus den Forschungsergebnissen entwickeln lassen?

CM: Hier geht es um die ganz konkrete Anwendung der In-Memory-Technologie. In Deutschland gibt es Krebsregister, in denen Daten zu Krankheitsverläufen für jeden Patienten erhoben und gespeichert werden. Wenn ich als Arzt konkrete, individuelle Patientendaten habe, also weiß wie alt mein Patient ist, wie groß sein Tumor ist und wo der angesiedelt ist und auch wie der bisherige Krankheitsverlauf war, kann ich nun dank der In-Memory-Technologie unmittelbar am Krankenbett, sekundenschnell Vergleichsdaten und Vergleichsabläufe über mein Tablett PC abrufen. Das gibt dem Arzt die Möglichkeit, aus den Daten des Registers erfahrungsbasierte Prognosen abzuleiten und entsprechende Therapien zu veranlassen.

Das, was hier in der Medizin durch die zentral gespeicherten Daten möglich ist, lässt sich aber auch auf andere Welten übertragen, etwa für die Durchsuchung eines umfänglichen Videoarchivs. So kann für den Nachruf für einen Politiker in Sekundenschnelle das Bildmaterial von Jahrzehnten erfolgreich durchsucht werden. Am Ende geht es immer wieder um den Zeitgewinn bei der Massendatenanalyse, und zwar bei konkreten, nicht standardisierten Abfragen.

Es hat in der Entwicklung der IT immer wieder Quantensprünge gegeben, bislang waren die Datenbanken an der Stelle eher ein Stiefkind. Die In-Memory-Technologie halte ich in der Tat für einen Quantensprung bei der Datenbanktechnologie. Es ist zum ersten Mal wirklich gelungen, das Thema Geschwindigkeit erfolgreich anzugehen und so die Analyse von Massendaten in Echtzeit zu ermöglichen.

Zur Person:

Prof. Dr. Christoph Meinel, Institutsdirektor und CEO der Hasso-Plattner-Institut GmbH

Internet-Technologien und Systeme stehen im Zentrum von Forschung und Lehre in dem von Prof. Dr. Christoph Meinel geleiteten Lehrstuhl. Er und sein Team aus wissenschaftlichen Mitarbeitern, Doktoranden und Studenten befassen sich mit der Erforschung und Entwicklung wissenschaftlicher Prinzipien, Methoden und Technologien zum Entwurf und zur Implementierung von Internet Technologien der Zukunft.