In-Memory-Datenbanken

Mehr Performance für Data Analytics

Datenanalysen ermöglichen datenbasierte Entscheidungen. Bei rasant wachsenden Datenmengen stossen diskbasierte Systeme an ihre Grenzen. In-Memory-Datenbanken sind eine performante Alternative.

→ VON JENS GRAUPMANN

DER AUTOR

Jens Graupmann
ist Vice President of Product Management von Exasol.
→ www.exasol.com

Big Data, Data Analytics, digitale Transformation – längst mehr als Schlagworte im unternehmerischen Alltag – verdanken ihre Bedeutung dem Umgang mit Daten. Massenhaft erfasst und in wertschaffende Zusammenhänge gebracht, bilden Daten je länger, je mehr die Basis für Geschäftsentscheidungen. Und das keineswegs durch ihre Aufbereitung im Rahmen wöchentlicher oder gar monatlicher Management-Reports. Vielmehr sind extrem zeitnahe Analysen gefragt, um interaktiv mit den Daten arbeiten zu können.

WENN PERFORMANCE ZÄHLT

Es geht also darum, sehr grosse Datenmengen in kurzer Zeit bearbeiten zu können. Klassische relationale Datenbanksysteme können bei solchen analytischen Anwendungsfällen zum Flaschenhals werden: Zum einen ist die Geschwindigkeit des Festplattenzugriffs limitiert. Zum anderen müssen dadurch immer grosse Datenblöcke eingelesen und blockweise verarbeitet werden. Doch insbesondere dann, wenn viele Datenmengen verknüpft werden sollen, dauert es lange, bis eine Auswertung über sämtliche Blöcke vorliegen kann. Der entstehende Overhead verlangsamt die Verarbeitung zusätzlich.

In-Memory-Datenbanksysteme (IMDBS) arbeiten anders: Sie nutzen den Hauptspeicher (RAM) als beschleunigenden Cache. Bis vor einiger Zeit trieb dieser Umstand die Kosten für ein IMDBS derart in die Höhe, dass solche Lösungen vor allem für spezielle, extrem datenabhängige Business Cases infrage kamen. Diese einst wenigen Fälle sind nun eher zum Alltag geworden: Immer mehr geschäftskritische Entscheide werden auf der Basis umfangreicher Datenanalysen getroffen. Zugleich sind die Preise für RAM stark gesunken, sodass die Dimensionierung des Hauptspeichers nicht stärker ins Gewicht fällt als die der anderen Komponenten. Um die vorhandene Hardware so gut wie möglich zu nutzen, wenden IMDBS die Prinzipien des Massive Parallel Processing (MPP) an. Auf dem Cluster-Level arbeiten sie nach dem SPMD-Paradigma (Single Programm, Multiple Data): Dabei führen mehrere Maschinen innerhalb des Clusters das gleiche Programm aus und verarbeiten eine Anfrage so gleichzeitig. Zusätzlich wird auf der Server-Ebene parallelisiert: Prozesse und Threads nutzen die leistungsstarken Funktionen moderner Multi-Core-Shared-Memory-Architekturen, was zu einer maximalen Auslastung der vorhandenen Standard-Server-Hardware führt.

Die IMDBS nutzen den Random Access auf die Daten besser aus. Jede Speicherzeile kann extrem schnell und direkt über ihre Speicheradresse angesprochen werden. Herkömmliche, diskbasierte Systeme haben hier das Nachsehen, selbst wenn sich der Datenblock im Cache der Datenbank befindet, da die verarbeitenden Algorithmen nicht für die Verarbeitung im Hauptspeicher optimiert sind.

«Immer mehr Geschäftsentscheide werden auf Basis von Datenanalysen getroffen»

Jens Graupmann

ALLES IM RAM, HOT DATA, ADD-ON

Auf dem Markt finden sich Datenbanklösungen mit unterschiedlich ausgeprägten In-Memory-Ansätzen. Im Wesentlichen lassen sich dabei drei Konzepte unterscheiden: IMDBS, die zur Datenverarbeitung ausschliesslich den Hauptspeicher nutzen, solche Lösungen, die nur die sogenannten Hot Data ins RAM laden und klassische Datenbanken, die eine In-Memory-Option als Add-on bieten. Die jeweilige Ausprägung entscheidet schliesslich darüber, wie viel kostspieliger Hauptspeicher benötigt wird und wie die übrige IT-Infrastruktur zu dimensionieren ist.

Bekanntester Vertreter der erstgenannten «Alles im Cache»-Variante ist SAP Hana. Alle Daten des Systems ste-

Bild: Shutterstock/Sergey Nivens

Daten stellen künftig die Grundlage für Geschäftsentscheide dar

hen permanent im Hauptspeicher zur Verfügung. So muss auch nichts von Festplatten oder anderen Laufwerken nachgeladen werden – ein eindeutiger Performance-Gewinn. Allerdings: Der Hauptspeicher muss entsprechend gross sein. Je mehr Daten verwendet werden – und für gewöhnlich wachsen Datenmengen im Laufe der Zeit eher progressiv als linear –, desto mehr RAM wird benötigt. Ist der Hauptspeicher voll belegt, können auch keine Daten mehr nachgeladen und verarbeitet werden. In der Praxis kommt es dabei zu Problemen im Ablauf, die erst durch die Beschaffung zusätzlichen RAMs beseitigt werden können. Darüber hinaus benötigen auch diese Systeme Festplatten, um die Persistenz sicherzustellen.

Die zweite IMDBS-Variante, vertreten beispielsweise durch Exasol aus Nürnberg, reduziert den Bedarf an Hauptspeicher, in dem sie – optimiert durch Algorithmen – nur die Daten im Speicher vorhält, die gerade verarbeitet werden. Diese «heissen Daten» machen etwa einen Drittel der Gesamt-Rohdaten-Menge aus, oft sogar erheblich weniger. Diese Kombination aus massiver Verarbeitung im Hauptspeicher und Datenablage auf anderen Speichermedien, wie etwa lokale Festplatten oder SAN-Speicher, kann dabei gleichzeitig Lastspitzen ohne Erweiterung besser abfedern. Allerdings müssen bei der Bearbeitung immer wieder Daten aus den angeschlossenen Speichermedien nachgeladen werden. In der Praxis erweisen sich die dadurch entstehenden Performance-Einbussen als eher gering, da eben selten über alle Daten gleichzeitig Abfragen laufen.

Komfortabel wäre es, könnte man bestehende klassische Datenbanken durch eine In-Memory-Lösung erweitern. Und tatsächlich bieten fast alle grossen Datenbankhersteller, wie etwa IBM und Oracle, eine In-Memory-Add-on-Option für ihre Systeme. Auch hier wird der Hauptspeicher genutzt, um die Verarbeitung zu beschleunigen. Die grundsätzliche Arbeitsweise ändert sich jedoch nicht: Als Add-on konkurriert die In-Memory-Funktion mit der eigentlichen Datenbank um den Hauptspeicher, der Performance-mindernde Overhead wird trotzdem erzeugt. Für die In-Memory-Option fallen zusätzliche Lizenzkosten an und Hardware muss nachgerüstet werden.

KOMPLETT ODER ERWEITERUNG?

In der Praxis hängt es natürlich vom konkreten Anwendungsfall ab, ob die komplette Migration auf ein neues IMDBS notwendig ist. Zumeist erweist es sich als sinnvoll, bestehende Systeme zunächst um eine IMDBS zu ergänzen. Dabei sind Unternehmen keineswegs auf den Anbieter beschränkt, von dem sie ihr bestehendes System bezogen haben – grundsätzlich können die Datenbanken herstellerübergreifend integriert werden. Hilfreich ist es dann, wenn das neue System für jegliche Art der Infrastruktur – von On-Premises über Private Cloud bis Public Cloud – geeignet ist. Bei der Wahl des IMDBS sollte deshalb auf offene APIs ebenso geachtet werden wie auf die Integrierbarkeit zusätzlicher externer Quellen und Business-Intelligence-Werkzeuge. Eine erweiterbare Integrationsschicht kann dabei als flexibler Zwischen-Layer fungieren. Zudem sollte die Datenbank gängige Big-Data-Technologien wie zum Beispiel Hadoop sowie typische Data-Science-Languages wie Python oder R unterstützen.

Die Entwicklung hin zum datengetriebenen Unternehmen hat gerade erst begonnen. Dabei geht es heute längst nicht mehr ausschliesslich darum, dieselben Informationen wie bisher nur ein wenig schneller auszuwerten. Vielmehr geht es um riesige Datenmengen und komplexe Zusammenhänge, die zukünftig ganz neue Anwendungsfälle ermöglichen. Die Performance bei der Datenanalyse ist deshalb schon jetzt ein Schlüsselfaktor. ←