Analytischer Daten-Turbo

Im Zeitalter von „Big Data“ müssen Analyse-Tools immer größere Datenmengen verarbeiten, was zwangsläufig zu Performanceproblemen führt. **Abhilfe verspricht das In-Memory-Konzept**, bei dem die Daten komplett im Arbeitsspeicher des Rechners vorgehalten und somit deutlich schneller verarbeitet werden. Was noch vor wenigen Jahren allein aus Kostengründen undenkbar erschien, ist heute Stand der Technik.

Dr. Marcus Dill

Nicht zuletzt seit SAP in das Geschäft mit In-Memory-Lösungen eingestiegen ist und mit großem Werbeaufwand eine eigene Technologie namens *HANA* (High Performance Analytic Appliance) anbietet, haben viele Unternehmen erkannt, dass sich im Bereich der Geschäftsanwendungen, insbesondere bei analytischen Anwendungen (Business Intelligence, BI) aktuell ein Umbruch vollzieht.

Alle großen Softwarehersteller sind mittlerweile auf diesen Zug aufgesprungen. Kennzeichnend für einen dynamisch wachsenden jungen Markt wie den der In-Memory-Technologie ist aber auch die Existenz und der Erfolg von kleinen Spezialanbietern. Wie nachhaltig ist dieser neue Trend? Was bringt der Einsatz von In-Memory-Werkzeugen im Bereich Business Intelligence?

Durchblick trotz Nebelkerzen

Die BI-Welt hatte über viele Jahre eine vertraute Ordnung mit klar definierten Begriffen und sauber abgegrenzten Tool- und Anwendungsklassen. Deren Eigenschaften, Vor- und Nachteile wurden von den meisten Beobachtern ähnlich wahrgenommen. Das Aufkommen von In-Memory-Technologie wirbelt den BI-Markt aktuell massiv durcheinander. Viele neue Begriffe und technische Ansätze, vor allem aber auch die vielen Nebelkerzen aus den Marketingabteilungen der Softwarehersteller machen den Durchblick selbst für Experten schwer.

Streng genommen ist die In-Memory-Technologie alles andere als neu. Es gibt sie seit vielen Jahrzehnten. Tatsächlich ist sie sogar das bei Weitem populärste BI-Tool der Welt – *Microsoft Excel* – im eigentlichen Sinn ein In-Memory-Werkzeug: Die zu analysierenden Daten werden vollständig in den Arbeitsspeicher geladen und dort mit oftmals beeindruckender Geschwindigkeit bearbeitet.

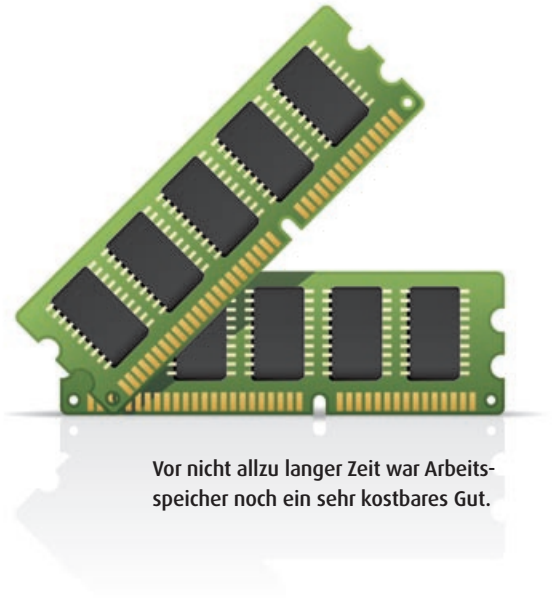
Auch wer Excel nicht als BI-Werkzeug akzeptieren mag, wird zugeben, dass beispielsweise mit *TM1* – auch heute noch eines der anerkannt besten Werkzeuge in seiner Anwendungsklasse – bereits in den 80er-Jahren ein In-Memory-Werkzeug für multidimensionale OLAP-Analysen (Online Analytical Processing) auf dem Markt war.

Ein Problem namens Festplatte

Dass sich In-Memory-Ansätze in der Vergangenheit nicht breit im Markt durchsetzen konnten, lag vor allem an der Limitierung im bearbeitbaren Datenvolumen. Arbeitsspeicher war lange Zeit eine sehr kostspielige Komponente von Rechnerarchitekturen. Daten im Unternehmensmaßstab im Arbeitsspeicher halten zu wollen, wäre bis vor wenigen Jahren schlicht unbezahlbar gewesen. Aber Geld war nur ein Grund: Erst die heute üblichen 64-Bit-Architekturen erlauben überhaupt das Adressieren eines entsprechend großen Speichervolumens.

Aus diesem Grund galt die Haltung großer Datenmengen innerhalb von Datenbanken oder in einem Dateiformat

Vor allem komplexe Analyseverfahren wie Data Mining profitieren vom In-Memory-Ansatz.



Vor nicht allzu langer Zeit war Arbeitsspeicher noch ein sehr kostbares Gut.

auf der Festplatte als unumgänglich. Entsprechend hatten manche Anbieter, die ursprünglich die arbeitsspeicherbasierte Technologie anboten, sogar nachträglich festplattenbasierte Datenablagen einschließlich einer entsprechenden Zugriffslogik geschaffen. Die Skalierbarkeit auch für große Datenmengen wurde seitdem mit Performance-Problemen erkauft, die in den meisten Fällen entweder innerhalb der Festplattenablage selbst oder im Datenzugriff und der Datenübertragung ihre Ursache haben.

Um überhaupt erträgliches Arbeiten mit großen Datenmengen zu ermöglichen, mussten sich Softwarehersteller und Anwenderunternehmen eine Vielzahl an Optimierungen und Workarounds einfallen lassen. Beispiele hierfür sind intelligentes Caching, spezialisierte Datenstrukturen wie OLAP-Cubes, aber vor allem Aggregate und andere Formen der redundanten Datenhaltung.

Optimierungen und Workarounds dieser Art hatten jedoch unerwünschte Nebeneffekte, die über ein komplexes Wirkungsgefüge zu hohen Kosten und geringer Nutzung eines BI-Systems führen. In nicht wenigen Fällen sind beziehungsweise waren diese Nebeneffekten die Ursache des Scheiterns oder des Dahinsiechens einer BI-Initiative.

Zunächst erfordern Entwicklung und Aufbau von Workarounds und die Optimierung von Datenmodellen großen Aufwand. Die Kosten hierfür übersteigen diejenigen für Softwarelizenzen in der Regel erheblich. Dass vielen Unternehmen internes Know-how und die Kapazitäten fehlen, verstärkt diesen Effekt, da man sich teure Experten und Berater von außen hinzuholen muss.

Oft sind die initialen Kosten für den Aufbau einer BI-Lösung aber noch gar nicht der Kern des Problems. Denn die Lösungen und Workarounds sind oft sehr komplex und kompliziert, was Betrieb und Wartung eines Systems auf Dauer sehr kostspielig machen können.

Schwerfällige Anwendungen

Komplexität und Redundanzen sorgen aber in vielen Fällen auch für Qualitätsprobleme: Liegen Daten mehrfach im System vor, besteht das Risiko von Inkonsistenzen in der Datenübertragung. In schlecht gemachten, vor allem aber in historisch gewachsenen Lösungen sind auch Logiken für die Datenverarbeitung redundant und leicht unterschiedlich implementiert. Sie werden schlicht zur Verarbeitung der an verschiedenen Stellen im System liegenden Daten mehrfach benötigt.

Eine nachträgliche Kapselung der Primär-Implementierung unterbleibt aus Zeit- und Aufwandsgründen, manchmal aber auch schlicht, weil diese undokumentiert und nicht allen Beteiligten bekannt ist. Auch wenn im BI-Umfeld gerne und zu Recht auf die aus den operativen Systemen ererbten Datenqualitätsprobleme verwiesen wird, tragen somit auch hausgemachte Fehler innerhalb der BI-Lösung zu einer eingeschränkten Verwendbarkeit und Verlässlichkeit der Daten bei.

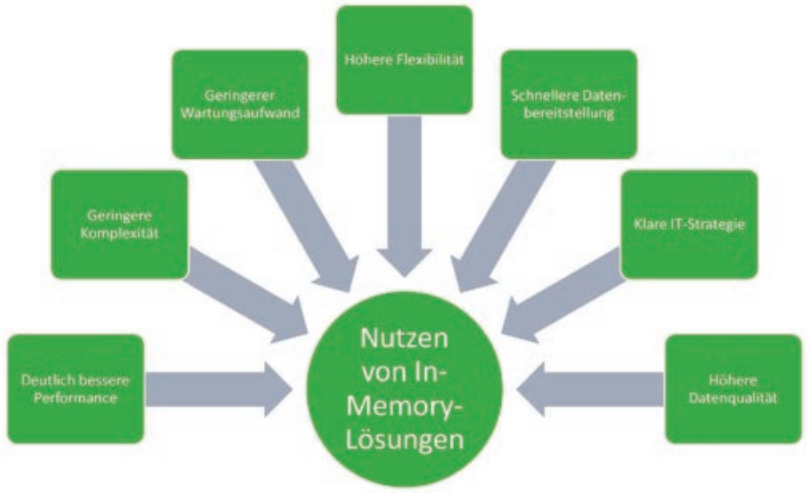
Neben diesen Qualitätsproblemen ist die geringe Flexibilität eine Hauptursache für die Unzufriedenheit mit BI-Lösungen. Auch dies lässt sich in vielen Fällen auf das Grundproblem der festplattenbasierten Datenablage in klassischen BI-Architekturen zurückführen. Datenwürfel, Aggregate und andere Datenstrukturen sind in den allermeisten Fällen auf vorher vereinbarte Anwendungszwecke hin optimiert.

Ändern sich die fachlichen Anforderungen, etwa durch den Bedarf der Analyse weiterer Daten, so zieht dies Erweiterungsarbeiten nach sich. Neben den Kosten hierfür ist oft die Geschwindigkeit, mit der diese Arbeiten erfolgen können, ein massives Problem.

Da in vielen BI-Landschaften die IT-Ressourcen bereits auf Monate verplant sind und eine zeitnahe Umsetzung von Anforderungen nicht möglich ist, werden die Fachbereiche auf diese Weise sehr häufig in Alternativlösungen gezwungen, um ihre kurzfristigen Analysebedarfe zu decken. Dieser Umstand trägt ganz wesentlich zum Wildwuchs klassischer BI-Lösungen bei, der in vielen Unternehmen zu beobachten ist.

In-Memory: mehr als Performance

Einige der Grundprobleme klassischer BI-Architekturen wurzeln also in der Aufbewahrung von Daten auf Festplatten und den damit verbundenen Performance-Engpässen. An dieser Stelle anzusetzen löst somit nicht nur die Performanceprobleme, sondern bereitet auch die Basis für die grundsätzliche Vereinfachung von System und Anwendungen. All die Optimierungen und Workarounds, die bisher unverzichtbar waren, um überhaupt mit einem System arbeiten zu können, werden durch den Einsatz von In-Memory-Architekturen überflüssig.



Der Nutzen der In-Memory-Technologie besteht nicht allein im Performance-Gewinn.

Letztendlich bedeutet die Einführung von In-Memory-Lösungen also weit mehr als eine hohe Performance. Sie führt vor allem auch zu einer deutlichen Kostenersparnis, einer stringenten IT-Strategie und vor allem zu zufriedenen, aktiven Anwendern, die den eigentlichen Return on Investment aus Daten und Systemen erst generieren.

Auch wenn die meisten In-Memory-Lösungen im Sinne der obigen Betrachtungen letztlich eine sinnvolle Investition sein können, lohnt der Vergleich verschiedener Angebote im Bereich der In-Memory-Analytik. Schließlich sind die Technologien im Markt sehr unterschiedlich leistungsfähig. Vor allem aber bringen sie eine Vielzahl an hilfreichen, ergänzenden Features mit sich.

Der Markt bietet mittlerweile eine Vielzahl von Produkten, die den verfügbaren Arbeitsspeicher teilweise ganz unterschiedlich verwenden und mit anderen technologischen Kniffen verbinden. Komprimierungstechniken beispielsweise können sehr unterschiedlich ausgeprägt sein. Eine starke Kompression der Daten hilft, mit dem verfügbaren Arbeitsspeicher effizient umzugehen.

Unterschiedliche Ansätze

Da die meisten Lizenzmodelle über das komprimierte Datenvolumen abgebildet sind, ist auch diesbezüglich die Kompression ein ganz entscheidender Faktor. Eine starke Kompression geht aber typischerweise einher mit Einschränkungen beim Update von Daten – nicht alle Hersteller bieten geeignete Deltaverfahren – und bei der Zugriffsgeschwindigkeit für bestimmte Arten von Abfragen.

Natürlich muss sich auch ein In-Memory-Werkzeug daran messen lassen, wie einfach sich Daten dort integrieren und organisieren lassen. Die Verfügbarkeit von Schnittstellen zu sozialen Netzwerken, Hadoop und anderen Big-Data-Datenquellen sollte hier beispielsweise betrachtet werden.

Auch die Modellierungswerkzeuge für den Aufbau der Datenstrukturen innerhalb des In-Memory-Tools unterscheiden sich deutlich. *QlikView* beispielsweise ermittelt Assoziationen zwischen verschiedenen Datentabellen automatisch und macht – zumindest theoretisch – eine explizite Modellierung überflüssig. Solche Automatismen sind aber nur so gut wie die impliziten Metadaten der jeweiligen Quellen. Folglich zwingt die Verwendung von *QlikView* zu einer gewissen Systematik und Disziplin bei der Bereitstellung und beim Import der Daten.

Ein weiteres Unterscheidungsmerkmal für In-Memory-Werkzeuge liegt in der Verfügbarkeit von komplexen Analyseverfahren (Data Mining, Predictive Analytics). SAS als langjähriger Marktführer in diesem Segment bietet seit einigen Monaten mit *SAS High Performance Analytics* eine Lösung, die umfassende Analytik auch für Big Data ermöglicht.

Aber auch SAP versucht, sich in diesem Umfeld zu profilieren. *SAP HANA* enthält mittlerweile auch eine Reihe von Algorithmen in einer sogenannten *Predictive Analytics Library*. Zusätzlich besteht die Möglichkeit der Integration von ***R***, der mittlerweile umfassendsten Data-Mining-Plattform. Nebenbei bietet *R* auch eine Vielzahl an Datenschnittstellen zu den wichtigsten modernen Datenquellen, allen voran sozialen Netzwerken. ***[rm]***

DER AUTOR

Dr. Marcus Dill
Geschäftsführer der mayato GmbH