Analytische Datenbanken – Hochleistungsmotoren für Business Intelligence

Es gibt sie schon seit über 15 Jahren: Neue Datenbanktechnologien, die darauf ausgelegt sind, riesige Datenbestände bei gleichzeitig hoher Anzahl von Abfragen durch viele bis sehr viele Nutzer in Sekundenschnelle zu analysieren. Aber erst heute finden sie Beachtung und eine stark zunehmende Nachfrage. Der Einsatz solcher analytischen Datenbanken, wie sie jetzt genannt werden, nimmt zu. Wir sehen hier als Trend einen weiteren Anstieg der Nachfrage nach analytischen Datenbanken in 2011, nicht nur ganz allgemein in Business Intelligence, sondern auch ganz besonders im analytischen CRM, wenn es um den Kunden und das Kundenwissen im analytischen CRM geht.

Der Treiber für analytische Datenbanken liegt auf der Hand. Das Datenvolumen steigt schneller als die Leistung von traditionellen Datenbanken. Man schafft es einfach nicht mehr, Daten im Detail zu analysieren, da es schlichtweg gesagt zu lange dauert. Gartner sagt in seinem Bericht zum Magic Quadrat for Data Warehouse Database Management Systems 2010: „Gartner-Klienten stehen bei der Abfrage von Data Warehouses immer häufiger vor erheblichen Performanceproblemen. Auf Grundlage dieser Informationen dürften rund 70 % aller Data Warehouses mit derartigen Problemen zu kämpfen haben."

Was machen analytische Datenbanken anders als herkömmliche Datenbanken? Da gibt es in der Tat verschiedene Methoden, die sich auch miteinander kombinieren lassen. Beginnen wir mit spaltenorientierten Datenbanken. Herkömmliche relationale Datenbanken sind zeilenorientiert. Das schafft bei großen Datenmengen einige Probleme, die wir jetzt zuerst beleuchten und danach die Vorteile von spaltenorientierten Datenbanken herausarbeiten.

Ein Datensatz, der Beispielsweise einen Kunden beschreibt, hat vielleicht sagen wir 1.000 Attribute, aber wir haben so viele Sätze, wie wir Kunden haben, also durchaus Millionen Sätze und vielleicht sogar noch mehr. Wenn wir nun in einer herkömmlichen Datenbank nach gewissen Kunden mit bestimmten Merkmalen (definiert über die Attribute) suchen, dann muss man eben alle Datensätze lesen. Beim Lesen stößt man gleich an ein ganz allgemeines Problem von herkömmlichen Datenbanken. Die sind nämlich gar nicht zum Lesen vieler Datensätze gebaut, sondern vom Design her eher transaktions-orientiert. Sprich, eine Datenbank gibt mir über einen Index in Bruchteilen von Sekunden eine bestimmte Datenmenge zum Ändern, Löschen oder Neuanlegen[1].

Will man also Adhoc-Abfragen auf herkömmlichen relationalen Datenbanken durchführen, dann braucht man Indizes und Aggregate, um schnelle Antworten zu erzielen. Das bedeutet aber, dass die Abfragen schon vorher bekannt sein müssen und durch Datenbankspezialisten aus der IT vorbereitet werden müssen (Sie bauen die Indizes und Aggregate). Mit anderen Worten, das ist teuer, weil gut bezahlte Spezialisten notwendig sind. Das ist zudem langsam: Denn wenn man mit einer neuen Idee kommt, zu der es noch keine Indizes und Aggregate gibt, dann müssen die erst gebaut werden. Wenn man eine Abfrage ohne eine solche Vorbereitung startet, kann der ganze IT-Betrieb empfindlich gestört werden. Indizes und Aggregate haben noch eine weitere unangenehme Eigenschaft: Sie brauchen Platz und machen die Datenbank um einen meist zweistelligen Faktor grösser als notwendig. Damit wird sie dann immer langsamer. Das führt dazu, dass irgendwann der Augenblick kommt, ab dem man gar keine Abfragen an die Datenbank mehr stellt, weil die Antworten viel zu spät eintreffen. Der Nutzer ist frustriert, das Wissen liegt brach in der Datenbank. Information wird zu einem reinen Kostenfaktor. Wissen über Kunden, Markt, Mitbewerber und Risiken lässt sich nicht mehr anwenden. An dieser Stelle stehen heute viele Unternehmen.

Analytische Datenbanken schaffen hier Abhilfe durch ihre Spaltenorientierung. Bei einer spaltenorientierten Datenbank kann jede Spalte in einer eigenen Datei liegen, d.h. auf einen Wert eines Attributs eines Datensatzes folgt in Lese-Reihenfolge nicht das nächste Attribut des selben Datensatzes, sondern das gleiche Attribut des nächsten Datensatzes: Die Zeilen und Spalten der Tabelle werden miteinander vertauscht. Intuitiv funktioniert dies, da in der Analytik meistens wenige Attribute von sehr vielen Datensätzen benötigt werden. Aufgrund der Spaltenorientierung müssen die restlichen Attribute nicht gelesen werden. Mit anderen Worten: das Lesen wird drastisch reduziert, weil man durch das Vertauschen von Zeilen und Spalten nur noch höchstens so viele Datensätze wie Attribute hat. Da die Anzahl der Attribute in der Regel klein ist gegen die Anzahl der Datensätze, bringt das einen hohen Performance-Gewinn. Jedoch wird das Schreiben von Datensätzen dadurch jedoch sehr teuer, was man aber oft durch Differenzdateien zum Teil ausgleichen kann.

Aufgrund dieser Basiseigenschaft von spaltenorientierten Datenbanken erhält man einen weiteren Vorteil. Man braucht keine Indizes und Aggregate mehr. Das macht die Datenbank schlanker, was wiederum das Lesen beschleunigt. Zusätzlich lassen sich die Daten dann komprimieren. Dazu werden einfache Verfahren genutzt, die es erlauben, relationale Operationen auf den komprimierten Daten auszuführen. So können beispielsweise mehrfach vorkommende Werte durch Kürzel fixer oder variabler Länge ersetzt werden, die durch ein Wörterbuch bei Bedarf wieder in die ursprünglichen Werte übersetzt werden können. Folgen identische Werte direkt aufeinander, können diese Sequenzen lauflängencodiert abgelegt werden. Sortierte ganzzahlige Daten können durch Differenzbildung zum jeweiligen Vorgänger oder zu einem lokalen Minimum in wenigen Bits untergebracht werden. Ein solches Komprimieren bringt also Kostenvorteile, da die Datenbank „klein" wird (Relativ zu einer zeilenorientierten Datenbank können die Daten bis zu 80% komprimiert werden.) Man erhält so weitere Performance-Vorteile.

Noch mehr Beschleunigung bringen neue Methoden, um auf komprimierten Spalten zu operieren. Die Benutzerschnittstelle bleibt zwar weiterhin SQL, aber die dahinter liegenden Zugriffsmethoden und -Algorithmen ändern sich. Das wollen wir hier nicht im Einzelnen diskutieren. Als Beispiele seien hier nur das parallele Scannen mehrerer Spalten und das von Google patentierte „MapReduce"-Verfahren genannt.

Eine weitere Beschleunigung lässt sich durch Parallelisieren der Verarbeitung auf Clustern und durch In-Memory-Verarbeitung erreichen. Das gilt sowohl für zeilen- wie auch spalten-orientierte Datenbanken. Daten werden dabei automatisch und gleichmäßig über alle Server eines Clusters verteilt, so dass für Abfragen alle Hardware-Ressourcen optimal ausgenutzt werden. Die Software ist so konzipiert, dass jeglicher Tuningaufwand entfällt, wie er in konventionellen Systemen üblich ist. Die Datenbanklösung legt Indizes automatisch an, analysiert und komprimiert die Daten selbständig und verteilt sie optimal über die Knoten. Intelligente Algorithmen fangen Server-Ausfälle auf und sorgen dafür, dass das System für Nutzer innerhalb weniger Sekunden ohne dessen Zutun wieder zur Verfügung steht.

Analytische Datenbanken werden in unterschiedlichen Ausprägungsformen angeboten. Es gibt parallelisierte herkömmliche Datenbanken, die in der Regel als Appliance angeboten werden, also eine spezielle Hardware und den parallelen Zugriffsmethoden und Algorithmen. Dabei sind solche Datenbanken dann immer noch zeilenorientiert.
Beispiele: EMC Greenplum, HP Neoview, Kognitio, IBM DB2 UDB DPF, IBM Netezza, Oracle Exadata, Teradata

Dann gibt es analytische Datenbanken, die spaltenorientiert sind, aber weitgehend Hardware-unabhängig eingesetzt werden können.
Beispiele: Apache Hadoop HBase, Illuminate, InfoBright, SAP Sybase IQ, Vertica

Und schließlich gibt es spaltenorientierte Datenbanken, die als Appliance teilweise mit spezieller Hardware angeboten werden, aber insbesondere In-Memory einsetzen.
Beispiele: Exasol, IBM Smart Analytics Engine, ParAccel, SAP HANA

Daneben gibt es auch noch besondere Verfahren wie beispielsweise „Database Images", das von Panoratio eingesetzt wird. Solche besonderen Verfahren bringen ähnliche Performanz- und Skalierungsgewinne.

Analytische Datenbanken lösen die Probleme, mit denen die Kunden heute kämpfen: Performance, Skalierbarkeit und Kosten. Fassen wir auch nochmal die Vorteile zusammen:

- Informationen sind flexibler abrufbar und stehen bis zu 100mal schneller zur Verfügung.
- Die Nutzerzufriedenheit erhöht sich signifikant aufgrund des schnelleren und flexibleren Zugriffs auf Information. Es können jetzt Daten analysiert werden, die vorher ohne Nutzen, aber mit Kosten gespeichert wurden. Das unterstützt und schafft bessere Entscheidungen.
- Die IT wird entlastet, da die analytischen Datenbanken hoch automatisiert sind und ein spezielles Wissen über Datenbankdesign und Tuning deutlich weniger gefragt ist.

Fallbeispiel[2]: ExaSolution, die Hochleistungsdatenbank von Exasol, bildet das neue Herzstück im Data Warehouse von xplosion. Der Anbieter von maßgeschneiderten Retargeting-Services für E-Commerce-Sites sichert sich mit ExaSolution einen nachhaltigen Wettbewerbsvorteil durch optimierte Prozesse. xretarget, das Kernprodukt von xplosion, ist eine weiterentwickelte Form des dynamisch personalisierten Retargetings. Dynamisch personalisiertes Retargeting basiert auf dem aktuellen Such- und Kaufverhalten von anonymen Profilen. Das Profil erhält ein Banner, das individuell um die Darstellung der von ihm zuvor angesehenen Produkte ergänzt wird. xretarget bietet einen entscheidenden Qualitätsunterschied: Data-Mining. Denn zeigt man dem Profil nur diejenigen Artikel an, die sein aktuelles Such- und Kaufverhalten widerspiegeln, führt das zu Wiederholungen: Immer wieder gleiche oder ähnliche Produkte im unspezifischen Banner. Bei der Ermittlung der passenden Produktempfehlung mit Hilfe des Data-Mining fließen die Informationen aus dem Such- und Kaufverhalten aller Profile ein. Statistische Analysen kategorisieren auf Basis übereinstimmender Eigenschaften neue Zielgruppen. Auf das einzelne Profil betrachtet bedeutet das: Es werden für das Profil Interessensfelder erschlossen, die es im Rahmen der Profilierung seines Such- und Kaufverhaltens nicht 1:1 gezeigt hat. Das Data Mining löst also die Fixierung auf vorgegebene Shop-Kategorien und macht den Blick frei für neues Verkaufspotenzial. So erhält jedes Profil das Banner, das optimal in Inhalt und Optik auf ihn zugeschnitten und dazu auch abwechslungsreich ist. Die Daten die hierfür nötig sind, werden im Data Warehouse von xplosion gesammelt und ausgewertet. Derzeit operiert das Unternehmen, das zur EOS Gruppe gehört und somit Teil der international agierenden Otto Group ist, mit Daten von mehreren Millionen anonymen Retargeting-Profilen. Auf Basis der Hochleistungsdatenbank von Exasol bietet xplosion seinen Kunden zukünftig schnellere Analysen, mit deren Hilfe nicht nur Retargeting-Kampagnen stetig optimiert werden, sondern das E-Business insgesamt erfolgreicher gestaltet werden kann.

Wir sehen für analytische Datenbanken in 2011 ein gutes Wachstum im gesamten Business Intelligence Umfeld. Nicht nur Handel, Konsumgüterhersteller und Versorger sitzen auf riesigen Datenmengen, sondern beispielsweise auch das Gesundheits- und Finanzwesen. Eines der Haupt-Treiber in 2011 wird sicherlich auch das zu erschließende Datenpotenzial in den Social Media" sein.

Zwei Dinge sollten zum Schluss noch klar gesagt werden:

- Eine analytische Datenbank macht ein physikalisches Datenbankdesign und Tuning weitgehend obsolet, aber sie ersetzt keineswegs das logische, fachliche Design der analytischen Datenbank. In diesem Sinne bleibt weiterhin ein Information Management unabdinglich, auch wenn analytische Datenbanken eingesetzt werden. Denn ein Stamm- und Metadaten-Management, ein Datenqualitäts-Management, eine Information Governance und die anderen Aufgaben im Information Management bleiben auch mit analytischen Datenbanken kritische Erfolgsfaktoren.
- Eine analytische Datenbank ersetzt nicht die herkömmlichen Datenbanken in der Transaktionsverarbeitung. Analytische Datenbanken sind eine neue Generation von Datenbanken für analytische Aufgaben im Unternehmen. Ein Unternehmen braucht heute eben zwei unterschiedliche Datenbanktechnologien, eine für die analytischen Aufgaben, eine für die Transaktionsverarbeitung.

Fazit: Analytische Datenbanken bringen den Nutzern ganz neue Möglichkeiten, sowohl in der Skalierbarkeit, der Performance als auch in den Betriebskosten. Neueste Techniktrends sind hierbei die Spaltenorientierung inklusive Komprimierung und Zugriffsverfahren, die massiv parallele Verarbeitung sowie die In-Memory-Technologie. Wer heute komplexe Analysen auf großen Datenmengen durch viele Benutzer mit vielen Abfragen ausführt und eine hohe Performance und Skalierbarkeit bei einfacher Wartbarkeit im benötigt, sollte analytische Datenbanken auf jeden Fall berücksichtigen. Wir meinen: Eine Evaluation lohnt sich auf jeden Fall. Damit sollte man auf keinen Fall mehr warten!

Dr. Wolfgang Martin
Wolfgang Martin Team
E-Mail: info@wolfgang-martin-team.net

Quelle:
[1] Das ist das sogenannte CRUD-Prinzip: „create, read, update, delete".

[2] Pressemitteilung der Exasol AG vom 14.12.2010 (http://www.exasol.com/single-message-view.html?&tx_ttnews[tt_news]=143&cHash=41c8ba602eefeefb042da89c876d70ca)