



Massive Parallel Processing als Turbo für Data Warehouses

Die Data-Warehouse-Systeme leiden in vielen Unternehmen unter technischen und ökonomischen Schwächen. Die Migration auf Massive Parallel Processing steigert die Performance und senkt die Komplexität.

KOMPAKT

- ▶ Klassische Data Warehouses scheitern bei Big Data
- ▶ Neue Architekturkonzepte erhöhen die Analyseleistung
- ▶ Data Scientists integrieren Daten aus sozialen Netzen

DIE EXPLOSION der Datenmengen bringt klassische Data Warehouses ins Schwitzen. Das Schlagwort Big Data bezeichnet riesige Mengen unstrukturierter Daten zum Beispiel aus Social Media und dem mobilen Internet. Nicht zuletzt diese Kanäle sind dafür verantwortlich, dass im vergangenen Jahr die weltweite Datenproduktion bereits auf über ein Zettabyte (das entspricht einer Milliarde Terabyte) angestiegen ist.

Gleichzeitig verlangen die Unternehmen tiefer gehende und schnellere Analysen von Daten, die stets verfügbar sein sollen. Immer mehr Mitarbeiter sollen diese Analysen nutzen. Kürzere Antwortzeiten und schnellere geschäftliche Einblicke werden gefordert sowie die Möglichkeit, viele verschiedene Datentypen und -quellen zu verwenden.

Big Data als Paradigmenwechsel

Wenn die Stichworte Social Media oder mobiles Internet fallen, hält sich das Interesse der Unternehmen zunächst in Grenzen. Das ändert sich schlagartig, wenn klar wird, welch wertvoller Schatz in diesen weitgehend unstrukturierten Daten schlummert.

Viele Unternehmen stehen heute unter großem Druck. Budgets werden immer knapper, Umsätze wach-

sen langsamer, stagnieren oder gehen zurück, gleichzeitig steigt die Dynamik der Märkte und damit der Wettbewerbsdruck stetig an. In einem solchen Umfeld können nur Unternehmen erfolgreich sein, die proaktiv zukunftsorientiert handeln und die folgenden Maßnahmen zur Chefsache erklären:

- Aufspüren von Umsatzpotentialen und Kosteneinsparungen,
- Intensivierung der Kundenpflege,
- effektiver Einsatz von Unternehmensressourcen.

Diese Maßnahmen sind überlebenswichtig. Dabei ist es stets von Vorteil, Entscheidungen mit Daten zu untermauern und dadurch mehr Sicherheit zu erlangen. Die Daten aus bestehenden Data-Warehouse-Systemen, die als Grundlage für Entscheidungen dienen, sind jedoch in der Regel retrospektiv. Auf

- In-Memory-Technologie krepelt bestehende Data-Warehouse-Konzepte um <http://tiny-cc/IM-Architektur>

der Basis von Vergangenheitsdaten Entscheidungen für die Zukunft zu treffen, kann jedoch nur gelingen, wenn die Rahmenbedingungen und das Umfeld des Unternehmens immer in etwa gleich bleiben. Diese Annahme ist heute unwahrscheinlicher als jemals zuvor.

Es steht außer Frage, dass Unternehmen ihre mit viel Aufwand etablierten Unternehmensprozesse weiterhin steuern müssen. Das reicht aber künftig nicht mehr aus. Unternehmen, die in Zukunft erfolgreich sein wollen, müssen willens und fähig sein, auf die Dynamik des Marktes zu reagieren. Das bedeutet, dass bestehende Prozesse angepasst und neue Abläufe etabliert werden müssen – einhergehend mit der Compliance, also der Erfüllung der gesetzlichen Vorgaben.

Genau an diesem Punkt kommt das Thema Big Data wieder ins Spiel. Kein Prozess funktioniert ohne Daten. Abläufe brauchen Daten und sie erzeugen Daten, die zur Steuerung dienen. Die Informationen, um aktuelle Trends und Entwicklungen zu identifizieren oder den Support und damit die Kundenbindung zu verbessern, finden sich massenweise in unternehmensinternen oder externen Produkt- oder Supportforen. Diese Daten sind extrem wertvoll, da sie hoch aktuell sind und zusätzlich Meinung zu Konkurrenten und deren Produkten enthalten. Die Auswertung dieser Informationen versetzt Unternehmen in die Lage, genau die richtigen Features und Produkte für den Markt zu entwickeln oder Service und Support gezielt zu verbessern.

Ein gutes Beispiel für den nicht realisierten Nutzen dieser Daten bietet der Markt für Mobiltelefone. Dort gab es bis vor nicht allzu langer Zeit einen absoluten Marktführer, der über Jahre seine Anteile ausgebaut hat. In sehr kurzer Zeit sind die Marktanteile dieses Herstellers drastisch gesunken. Dieses Unternehmen hatte sich keine Mög-

lichkeit geschaffen, Trends sinnvoll auszuwerten und daher die Potenziale moderner Smartphones und die steigende Nachfrage danach nicht erkannt, die Technologieführerschaft verloren und dann zunächst auch noch auf den möglicherweise falschen Betriebssystem-Partner gesetzt.

Die Textsearch analysiert auch unstrukturierte Daten

Bleibt die Erkenntnis und Problematik, dass es sich bei Big Data in der Regel um unstrukturierte Daten handelt. Das Herausfiltern der

und reichern die bestehenden Data Warehouses damit an.

Die großen Datenmengen stellen nur einen Teilaspekt des Themas Big Data dar. Eng verbunden damit sind neue Data-Warehouse-Fähigkeiten, eine andere Einschätzung des Werts von Daten sowie der Aufbau einer Infrastruktur, die auf der Suche nach bisher unbekanntem Beziehungen zwischen Produkten, Kunden und Lieferanten wirklich alle Daten eines Unternehmens einbezieht.

So riesig wie die Datenmengen ist das Nutzenpotenzial für die Un-

Für Big-Data-Analysen müssen Unternehmen ihr Data-Warehouse-Konzept neu aufsetzen.

wichtigen Daten erfolgt mit Data Mining und Textsearch-Algorithmen. Für die Transformation in eine strukturierte, auswertbare Form kommen eine Reihe neuer Produkte zum Einsatz, die Hadoop-Technologie mit MapReduce verwenden – also automatisierte Verfahren, wie Anwendungen ihre Daten finden. In den USA hat sich in diesem Zusammenhang bereits ein neues Berufsbild entwickelt, der Data Scientist. Diese hochbezahlten Spezialisten filtern und strukturieren die Daten aus sozialen Netzwerken

Unternehmen. Es wird darauf ankommen, die technischen Hürden für deren umfassende und schnelle Analyse zu nehmen. Unternehmen sind gefordert, rechtzeitig die technischen Voraussetzungen zu schaffen für die neuen Herausforderungen im Data Warehousing, aber auch für andere, die sicher noch kommen werden. Dabei sollten sinnvollerweise die Betriebskosten gesenkt und die getätigten Investitionen in die Datenbeschaffungsprozesse und die Auswertungen geschützt werden.

Die Autoren



Wolfgang Dähler (links) und Heinrich Smielowski sind beide Geschäftsführer der DATA MART Consulting GmbH.

Genau hier schließt sich der Kreis: Mit der Verarbeitung dieser Massendaten und dem damit einhergehenden Zuwachs an Analysefähigkeit ist ein herkömmliches Data Warehouse überfordert. Der Umstieg auf Massive Parallel Processing ist die logische Konsequenz aus diesen Überlegungen. Ein Data Warehouse mit dieser Technologie skaliert sehr gut und verfügt über

genügend Leistungs- und Kapazitätsreserven für die nächsten drei bis fünf Jahre. Zudem wirkt sich diese Architektur positiv auf Kosten, Performance, Komplexität und Funktionalität aus.

Die Parallelisierung verbessert vier Bereiche

In Bezug auf Massive Parallel Processing findet derzeit ein massiver

Machtkampf und Verdrängungswettbewerb im Markt der großen Datenbankhersteller statt. Das große Interesse der Anwender an diesem Konzept ist nur zum Teil auf die intensiven Marketinganstrengungen der Anbieter zurückzuführen, die diese Technologie zum Standard für Data Warehousing machen wollen. Massive Parallel Processing bietet tatsächlich mehrere Vorteile bei der Ablösung bestehender Data-Warehouse-Systeme:

Kosten

Durch den harten Wettbewerb um Marktanteile sinken die Anschaffungspreise und auch die Wartungskosten für Massive Parallel Processing Appliances monatlich. Auch für mittelständische Unternehmen ist der Umstieg inzwischen durchaus reizvoll. Es müssen auch keine Heerscharen an Beratern mehr beschäftigt werden, die Hilfstabellen, Materialized Views, Bitmap-Indizes und SQL-Programme erstellen, um beispielsweise Ladeprozesse einige Minuten schneller zu machen. Auch die ständigen Bemühungen, auf Drängen der Anwender Reporting und Analyse zu beschleunigen, sind damit hinfällig. Stattdessen reduziert sich durch die automatischen Optimierungsmodelle des Master-Servers die gesamte Datenbank-Administration auf einen Bruchteil.

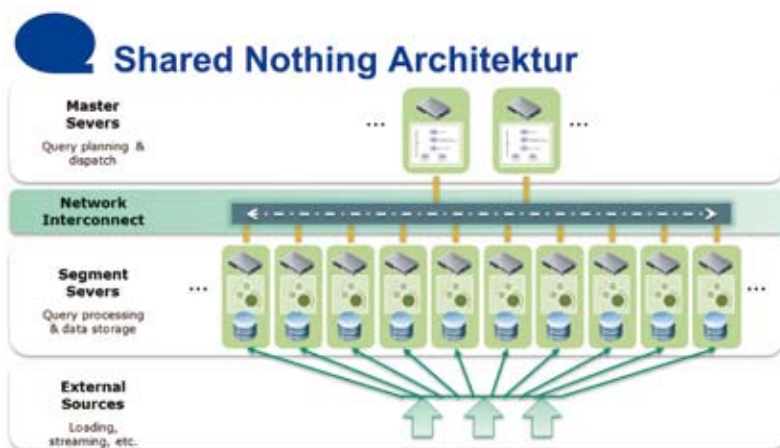
Performance

Sämtliche Anbieter von Massive Parallel Processing gehen je nach Aufgabenstellung von einem Performancegewinn mit Faktor 10 bis 120 aus. Das bringt in jedem Fall genügend Reserven für die mittlere Zukunft. Zusätzlich sind die neueren Systeme auch im laufenden Data-Warehouse-Betrieb umfangreich skalierbar.

Komplexität

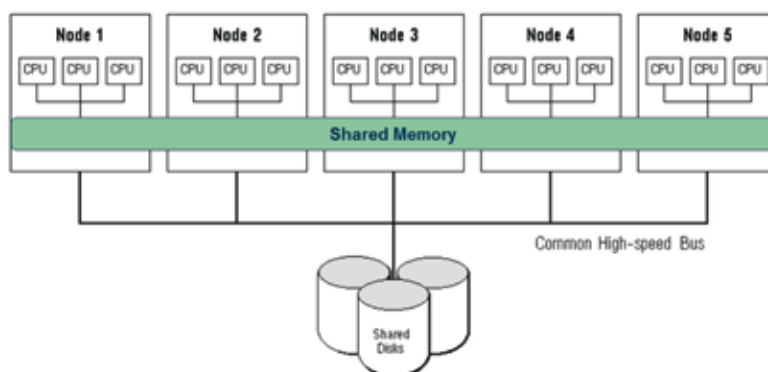
Die Komplexität des Datenbankschemas wird durch den Wegfall von Hilfstabellen, Materialized Views und Bitmap-Indizes erheb-

Beim Massive Parallel Processing stehen mehrere Architekturkonzepte zur Auswahl



In der Shared-Nothing-Architektur erhält jeder Knoten einen Teil der Gesamtdaten, die er mit eigenen Ressourcen bearbeitet. Ein sogenannter Master-Server organisiert sowohl bei ETL-Prozessen (Extraktion, Transformation, Laden) als auch bei abfrageorientierten Prozessen die optimale Verteilung der Daten und der Last auf die Knoten.

Shared Everything Architektur



Bei der Shared-Everything-Architektur teilen sich die Knoten alle verfügbaren Ressourcen und nutzen einen gemeinsamen Datenbestand, der meist auf Storage-Servern liegt. Bei reinen Data-Warehouse-spezifischen Aufgaben zeigen sich Performance-Nachteile, dafür eignet sich das Konzept auch für Online Transaction Processing.

Quelle: Data Mart Consulting

lich reduziert. Der größte Unterschied zum herkömmlichen Data Warehouse liegt darin, dass durch den großen Performancegewinn weitgehend auf die verdichtete Analyseschicht verzichtet werden kann. Der nächtliche Ladeprozess, bei dem die Verdichtung der atomaren Analysedaten einen großen Teil der Gesamtzeit in Anspruch nimmt, verkürzt sich somit deutlich.

Funktionalität

Massive Parallel Processing bringt einige Data-Warehouse-Funktionalitäten, die man bisher lediglich von mehrdimensionalen OLAP-Da-

konzept Performancenachteile. Es eignet sich aber im Gegensatz zur Shared-Nothing-Architektur auch für OLTP-Anwendungen (Online Transaction Processing).

Die Migration des Systems kann in drei Schritten erfolgen

Mit dem Komplettpaket *DWH-RetroFIT* bietet DataMart Consulting eine Methodik zum Umstieg auf ein System mit Massive Parallel Processing, die einen Transfer des bisherigen Data Warehouse auf die neue Plattform gewährleistet. Ziele sind hierbei die Steigerung der

Master-Server steuert Verteilung der Daten und Last beim Shared-Nothing-Konzept.

tenbanken (Online Analytical Processing) kannte. Verdichtungsfunktionen zur Laufzeit ermöglichen in den meisten Fällen den Verzicht auf Aggregate mit verdichteten Daten.

Shared Nothing versus Shared Everything als Alternativen

Beim Massive Parallel Processing unterscheidet man grundsätzlich zwischen zwei Architekturen: Shared Nothing und Shared Everything. In der Shared-Nothing-Architektur erhält jeder Knoten (Node) einen Teil der Gesamtdaten, die er mit eigenen Ressourcen bearbeitet. Ein sogenannter Master-Server organisiert sowohl bei ETL-Prozessen (Extraktion, Transformation, Laden) als auch bei abfrageorientierten Prozessen die optimale Verteilung der Daten und der Last auf die Knoten. Die Shared-Nothing-Architektur eignet sich ausschließlich für Data-Warehouse-Anwendungen und erzeugt hier sehr gute Performance.

Das Alternativkonzept stellt die Shared-Everything-Architektur dar. Hierbei teilen sich die Knoten alle verfügbaren Ressourcen und nutzen einen gemeinsamen Datenbestand, der meist auf Storage-Servern liegt. Bei reinen Data-Warehouse-spezifischen Aufgaben zeigt dieses

System-Performance, die Senkung der Betriebskosten und die Erweiterung der Kapazitäten unter der gleichzeitigen Beibehaltung der vorhandenen ETL-Prozesse (Extraktion, Transformation, Laden) und Business-Intelligence-Systeme, inklusive der bestehenden Auswertungen. Die Lösung besteht aus drei Bausteinen:

1. Health-Check

Eine grobe Untersuchung des bestehenden Data-Warehouse-Systems deckt Schwächen auf und prüft die Möglichkeit für ein *DWH-RetroFIT*.

2. Proof of Concept

Beim Auftraggeber werden vor Ort auf einem produktionsnahen Massive-Parallel-Processing-System Tests durchgeführt, die aufzeigen, wie Systemschwächen beseitigt und System-Performance und -Ökonomie verbessert werden können.

3. Migration

Bei der eigentlichen Ablösung des Systems wird die Massive-Parallel-Processing-Plattform produktiv gesetzt. Alle bestehenden Daten werden auf diese Plattform transferiert und alle notwendigen Schnittstellen werden umgestellt. Darüber hinaus umfasst die Migration Tests sowie die Übergabe an die betriebsführenden Organisationen. ◀