Hochschule für Technik, Wirtschaft und Kultur Leipzig

Fakultät Informatik und Medien

Oberseminar
„Datenbanksysteme – Aktuelle Trends“
Prof. Dr.-Ing. Thomas Kudraß, SS 2019

Abstract
Hauptspeicherdatenbanken

von
Carina Landerer
MIM-18
Matrikel-Nr.: 73060

Leipzig, 30.Juli 2019

1. Einleitung

Wachsende Datenmengen machen das *Bedürfnis* nach schneller Datenverarbeitung zu einer großen Herausforderung. Um die Menge an Daten, die in kürzester Zeit im Internet entstehen sinnvoll nutzen zu können, müssen diese Daten bestenfalls in Echtzeit analysiert werden.

Hauptspeicherdatenbanken bieten eine Lösung zu dieser komplexen Aufgabenstellung, indem Daten direkt aus dem Hauptspeicher geladen und von dort auch verarbeitet werden können. Daher kommt diese Technologie häufig in Geschäftsbereichen zur Anwendung, in denen eine solche Echtzeitanalyse und –Datenverarbeitung notwendig ist.

Erreicht wird dieser Vorteil hauptsächlich durch die Zusammenführung der historisch klar getrennten Bereiche OLTP und OLAP.

1.1 OLTP und OLAP

OLTP (Online Transaction Processing) ist stark transaktionsorientiert und steht in direkter Verbindung mit der Echtzeitverarbeitung von Transaktionen. Dabei stellt auch die Transaktionssicherheit eine zentrale Komponente dar. In OLTP Anwendungen werden zusammengehörige Operationen stets in Transaktionen zusammengefasst und durch vordefinierte Transaktionskriterien (ACID-Kriterien) gewährleistet, dass die Konsistenz der Datenbank zu jedem Zeitpunkt sichergestellt ist. Somit kann mit OLTP eine schnelle und direkte Verarbeitung von Daten stattfinden.

Bei OLAP (Online Analytical Processing) stehen hingegen nicht die Transaktionen sondern die komplexe Analyse von riesigen Datenmengen im Vordergrund. Das Wichtigste ist also die historische und aggregierte Information, weshalb der Zugriff meist nur lesend erfolgt. OLAP dient daher hauptsächlich der Entscheidungsfindung auf Basis von aufbereiteten Daten bei denen der Fokus mehr auf historischen Analyse als auf der Verarbeitung absolut aktueller Werte liegt.

Durch diese klare Aufgabentrennung lassen sich die Meisten Operationen gut in OLTP und OLAP einteilen und entsprechend verarbeiten. Allerdings gibt es auch Anwendungsfälle, in denen es nötig wird, komplexe Daten in Echtzeit abzufragen und an dieser Stelle stößt der Ansatz oft an seine Grenzen. [1]

1.2 Probleme in Geschäftsprozessen

Komplexe Analysedaten sind oftmals stark mit der Business Logik verbunden und daher sehr langsam abzufragen. Zwei typische Anwendungsfälle bei denen

eine Echtzeitanalyse trotzdem notwendig ist, sind der Available-to-Promise-Check und die Erstellung von Mahnungen.

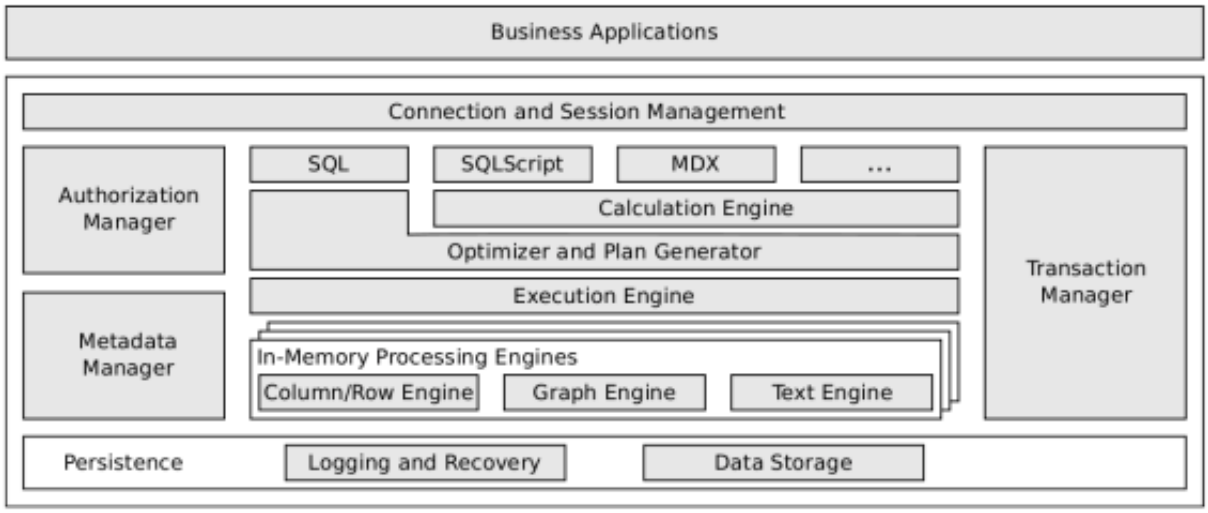
Das „Mahnungsproblem“ steht in direkter Verbindung mit SAP. In den SAP-Systemen ist es den Geschäftspartnern möglich, bei überfälligen Zahlungen automatisch Mahnungen an ihre Kunden zu generieren. Dabei müssen zunächst mehrere Parameter geprüft werden, bevor eine Mahnung automatisiert an den entsprechenden Kunden geschickt werden kann. Das ist ein komplexer Vorgang, welcher mit herkömmlichen Datenbanksystemen zu etwa 20 Minuten pro Mahnung geführt hat, während er bei der Verwendung von Hauptspeicherdatenbanken nur noch wenige Sekunden benötigt.

Ein anderer komplexer Vorgang ist der „Available-to-promise“-Check. Dieser kontrolliert die Verfügbarkeit einer bestimmten Menge eines Materials oder Produkts zu einem bestimmten Zeitpunkt und berechnet gegebenenfalls alternative Verfügbarkeitstermine oder –mengen. Dieser Check wird meist sowohl innerbetrieblich entlang der *Supply Chain* als auch im Vertrieb angewendet um möglichst schnell Engpässe zu identifizieren. Durch diesen unternehmensübergreifenden Kontext steigt die Komplexität der Berechnungen. Um dennoch eine Verarbeitung in nahezu Echtzeit zu gewährleisten, wird an dieser Stelle ebenfalls auf die Alternativen der Hauptspeicherdatenbanken zurückgegriffen. [6]

2. Hauptspeicherdatenbanken

Klassischerweise liegen in Datenbanksystemen die Daten auf externen Speichern und werden bei Verwendung direkt von dort angefragt. Das kommt historisch betrachtet zunächst einmal daher, dass der Arbeitsspeicher eines Computers zunächst sehr beschränkt war. Heute hat sich die Hardware allerdings stark verändert, sodass es nun möglich ist, komplette Datenbanken im Hauptspeicher zu hosten. Damit benötigen Hauptspeicherdatenbanken im Gegensatz zu herkömmlichen Datenbankmanagementsystemen keine Festplattenlaufwerke mehr zum Datenzugriff.
Das steigert vor allem die Leistung bei hohen Zugriffsraten oder bei der Echtzeitanalyse von Datenmengen wie beispielsweise Ereignisdaten, auf die möglichst schnell reagiert werden muss.
Zugrunde liegt die höhere Leistungsfähigkeit von Hautspeichern im Vergleich zu Festplatten und die Möglichkeit der Parallelität von Abfragen und Verarbeitungen. [2]

2.1 Architektur



Die erste Applikationsschicht unterscheidet sich bei Hauptspeicherdatenbanken nicht von den Herkömmlichen. Zunächst gibt es die Business Anwendung, in der die Kommunikation des Nutzers mit der Anwendung stattfindet und Queries erstellt werden.
Während allerdings in herkömmlichen Datenbanken nur die Logik im Hauptspeicher ausgeführt wird, wird diese Ebene in In-Memory-Datenbanken noch erweitert. In dieser Technologie wird die ganze Datenbank mit allen Daten im Hauptspeicher gehalten und nicht extern auf Festplatten ausgelagert.

Dennoch bleibt der Hauptspeicher ein flüchtiges Medium, weshalb auch diese Systeme nicht ganz ohne persistente Sekundärspeicher auskommen. Dort werden Backup- und Logging-Dateien gesichert um im Falles eines plötzlichen Absturzes der Anwendung die Transaktionssicherheit zu gewährleisten. Aus diesen Daten soll sich zu jedem Zeitpunkt ein aktueller Stand der Datenbank regenerieren lassen. [2]

3. SAP HANA

Obwohl mittlerweile viele großen Softwareanbieter wie Oracle Möglichkeiten zur Integration von Hauptspeicherdatenbanken bieten, bleibt SAP mit seiner selbst entwickelten Hauptspeicherdatenbank „SAP HANA“ der Vorreiter in der Technologie. [5]

3.1 Spaltenorientierte Datenorganisation

Während die meisten relationalen Datenbanksysteme ihre Daten zeilenorientiert abspeichern, werden die Daten in HANA spaltenorientiert abgespeichert. Das bedeutet, dass nicht die einzelnen Zeilen der Datensätze hintereinander auf dem linearen Speicher des Computers abgebildet werden, sondern immer spaltenweise hintereinander gespeichert wird.

Das führt zu deutlichen Performance-Vorteilen bei der Abfrage großer Datenmengen für OLTP Operationen, bei denen Scans vor allem über einige wenige Attribute durchgeführt werden. Das kommt daher, dass bei dieser Organisation nicht alle Attribute eines Datensatzes ausgelesen werden müssen, sondern nur die wirklich Relevanten.
Zudem sind mit dieser Datenstruktur einige Kompressionsmöglichkeiten wie die Wörterbuchkodierung oder Lauflängenkodierung möglich. [7]

3.2 Abfragetechniken

3.2.1 SELECT

Der Select Befehl funktioniert je nach Komprimierung ähnlich wie bei herkömmlichen Datenbanken. Wird wie in HANA eine Wörterbuchkodierung verwendet, so wird zunächst der Attribut-Vektor gescannt um die Position des Datensatzes in der eigentlichen Tabelle zu ermitteln. Anschließend wird mithilfe dieses Wertes auf den eigentlichen Datensatz zugegriffen. [3, S. 103-106]
Durch die spaltenorientierte Datenorganisation gibt es allerdings ein Problem bei der Abfrage „SELECT *“, da in diesem Fall alle einzelnen Spaltentabellen nach dazugehörigen Werten durchsucht werden müssen, was unter Umständen sehr lange dauern kann. Daher ist es von größter Bedeutung, die relevanten Daten immer direkt und explizit abzufragen. [7]

3.2.2 INSERT ONLY

Der INSERT Befehl kann bei einer spaltenorientierten Datenbank sehr schreibintensiv werden, da nicht nur an einer Stelle Daten hinzugefügt werden müssen, sondern für jede Spalte ein gesonderter Zugriff erfolgt. In HANA wurde dieses Problem über einen Deltaspeicher gelöst, welcher zeilenorientiert die Daten aufnimmt und anschießend in periodischen Abständen die Daten in die zeilenorientierte Hauptdatenbank überträgt.

Zudem kommt die INSERT-ONLY Strategie zum Einsatz. Das bedeutet, dass grundsätzlich nur Daten hinzugefügt werden, aber keine Datensätze tatsächlich gelöscht oder verändert werden. Das bedeutet allerdings nicht, dass keine logischen DELETE oder UPDATE Operationen durchgeführt werden können. Stattdessen wird allen eingefügten Daten ein logischer Marker hinzugefügt, der als Gültigkeitsflag dient. Damit bleiben alle hinzugefügten Daten nur in einem bestimmten Zeitraum gültig. Wenn nun auf den entsprechenden Datensatz neue Daten hinzugefügt werden, bekommen diese Daten einen aktuelleren Gültigkeitsflag, womit der Eintrag als geändert gilt. Beim Löschen wird dem neuen Datensatz ein Ungültigkeitsflag hinzugefügt, sodass ein logischer DELETE stattgefunden hat. [3, S.73-81]

3.2.3 Anfrageverarbeitung

Abfragetechniken betreffen vor allem alle Arten von Queries, die Daten abfragen und selektieren, dazu gehören vor allem Joins und Aggregationsfunktionen.

Es gibt verschiedene Arten von Joins, von denen die Meisten aus herkömmlichen Datenbanksystemen bereits bekannt sind. Im Bezug auf Hauptspeicherdatenbanken wie HANA gibt es vor allem zwei Arten von Joins, welche die Spaltenorientierung bestmöglich ausnutzen: der Hashjoin, welcher auf einer Hash-Funktion basiert und der Sort-Merge-Join. Für diese Version muss eine weitere Übersetzungstabelle eingeführt werden, welche WertIds der verbundenen Spalten aus beiden Tabellen beinhaltet. Somit entsteht zu jeder Tabelle zusätzlich zu ihrem eigenen Wörterbuch eine gemeinsame Übersetzungstabelle und ein Attribut-Vektor. [3, S.135-141]

Aggregationsfunktionen sind beispielsweise COUNT oder SUM. Grundsätzlich laufen auch diese Operationen nach einem ähnlichen Prinzip ab wie die beschriebenen Joins. Es wird aufgrund der Spaltenorientierung versucht, so sequentiell wie möglich vorzugehen. Daher wird zunächst der Attribut-Vektor der gesuchten Spalte durchlaufen und ein Zähler für jede neue WertId eingeführt. Beim Durchlaufen der eigentlichen Tabelle wird dieser Zähler nun bei jeder Übereinstimmung erhöht, womit am Ende eine Ergebnistabelle mit gezählten Werten erstellt werden kann. [3, S.145-147]

4. Zusammenfassung

Zusammenfassend lässt sich sagen, dass Hauptspeicherdatenbanken in den letzten Jahren stark an Bedeutung gewonnen haben. Zahlreiche Hersteller versuchen aktuell, ihre Daten näher an die Logik zu bringen und daher mehr Operationen im Hauptspeicher auszuführen.
Durch sinkende Hardware-Kosten wird diese Technik langsam auch für kleinere Unternehmen rentabler. [2] Besonders wichtig bleiben die Hauptpeicherdatenbanken allerdings im Big Data Umfeld, da sie dort entscheidende Verarbeitungsvorteile gegenüber herkömmlichen Datenbanksystemen mit sich bringen.

5. Literaturverzeichnis

[1] Computerwoche, Ad-Hoc- Analysen mit OLAP https://www.tecchannel.de/a/bi-methoden-teil-1-ad-hoc-analysen-mit-olap,1751285,2 , Zugriffsdatum: 10.05.2019

[2] Jens Lechtenbörger und Vanessa Jie Ling, Abschlussbericht zum Projekt „Das Potential von In-Memory-Datenbanken in und für KMU“, April 2015

[3] Plattner, Hasso. Lehrbuch In-Memory Data Management : Grundlagen Der In-Memory-Technologie. 2013. Print.

[4] Krueger, Jens, Martin Grund, Christian Tinnefeld, Benjamin Eckart, Alexander Zeier, and Hasso Plattner. "Hauptspeicherdatenbanken Für Unternehmensanwendungen." Datenbank-Spektrum 10.3 (2010): 143-58. Web.

[5] Chandrasekhar Mankala und Ganesh Mahadevan. „SAP HANA Cookbook“, 2013

[6] AKASH KUMAR , November 2015, „How to Plan for Disaster Recovery with SAP HANA”, https://blogs.sap.com/2015/11/13/how-to-plan-for-disaster-recovery-with-sap-hana , Zugriffsdatum: 11.05.2019

[7] ComputerWeekly.de, Spaltenorientierte Datenbank, https://www.computerweekly.com/de/definition/Spaltenorientierte-Datenbank, Zugriffsdatum: 11.05.2019