Aktuelle Trends aus Business Intelligence & Datawarehouse

Autor: Klaus Rohrmoser

Es entstehen immer größere Datenmengen, die in immer unterschiedlicheren Formaten und aus immer mehr Datenquellen gespeist werden. Die Erkenntnisse, die aus diesen Daten gewonnen werden können, sind das Gold des Informationszeitalters. Nicht umsonst können Soziale Medien wie Facebook oder Google Ihre Dienste kostenlos zur Verfügung stellen. Daten intelligent zu verwenden wird künftig immer stärker im Fokus stehen. Hierbei gibt es verschiedenste Methoden und Technologien, die unter den Begriffen Business Intelligence und Datawarehouse zusammengefasst werden können. Dieser Artikel gibt einen Überblick über aktuelle Trends im Business Intelligence und Datawarehouse Bereich.

Business Intelligence und Datawarehouse werden oft als Synonym verwendet, unterscheiden sich jedoch grundsätzlich. Während der Begriff Datawarehouse Technologien zur optimierten Datenspeicherung umfasst, ist Business Intelligence als ein Prozess zu verstehen, der relevante Informationen aus Daten gewinnt und diese an operative Systeme zurückspielen kann (Closed Loop). Business Intelligence kann IT Systemen und Anwendern ein "Lernen" aus verfügbaren Daten ermöglichen, um effizientere Entscheidungen zu treffen.

Business Intelligence - Self Service BI

In vielen Unternehmen ist die IT an vorgegebene Releasezyklen, einzuhaltende SLAs und Kostenlimits gebunden. Fachanwender möchten jedoch eine schnelle und dynamische Umsetzung Ihrer Anforderungen, um ihr Geschäft zu betreiben. Dieser Widerspruch lässt sich durch Self Service Business Intelligence auflösen, sofern einige wichtige Aspekte beachtet werden.

Mit Self Service BI kann mehr Agilität in der Analyse und Auswertung von Unternehmensdaten generiert werden. Reporting Anwender können Anforderungen schnell und flexibel mit den vorhandenen Reportingsystem(en) umsetzen, durch Ad Hoc Analysen und Reports. Damit können Unternehmen schneller auf Anforderungen von Kunden und dem Geschäftsumfeld reagieren.

Aspekte des self-service BI:

- *Benutzerfreundlichkeit*: einfache Bedienbarkeit des Reportingsystems bei der Erstellung von Ad Hoc Reports oder Analysen. Eine Möglichkeit zur Zusammenarbeit zwischen Benutzern (Collaboration). Integration von eigenen - oft Excelbasierten - Auswertungen in ein bestehendes Dashboard.
- *Rollenverständnis*: Fachanwender lieben Excel und können damit ausgefeilte Auswertungen erzeugen. Data Analysten sind SQL affin, verstehen Datenmodelle und können komplexe Analysen erstellen. Die IT muss eine Infrastruktur finden, um beiden Rollen einen Mehrwert bieten zu können.
- *Data Governance*[1]: Datenqualität, Datenherkunft und Aktualität, Kennzahldefinitionen sind entscheidende Aspekte, damit bei der Anwendung von Self Service BI keine Äpfel mit Birnen verglichen werden sondern verlässliche und nachweisbare Informationen erzeugt werden.
- *Agilität & Business Intelligence*: Ein ständiger Austausch zwischen Fachanwender, Data Analysten und IT um mit Hilfe der gefundenen Erkenntnissen IT-Systeme weiter zu entwickeln und/oder Geschäftsprozesse zu verbessern.
- *Gemeinsame Datenbasis*: Eine über mehrere Quellen integrierte Datenbasis für Reporting Stakeholder, die z.B. in einem Datawarehouse vorhanden ist. Flexibel auswertbar, um zusätzliche (z.B. Fachbereich bezogene) Datenquellen erweiterbar, unter Einhaltung der Data Governance.

1 "Data governance is a set of processes that ensures that important data assets are formally managed throughout the enterprise. Data governance ensures that data can be trusted and that people can be made accountable for any adverse event that happens because of low data quality. Quelle: http://en.wikipedia.org/wiki/Data_governance"

- *Sandboxing*[2]: Durch das einmalige Bereitstellung von produktiven Daten zur Daten- und Anforderungs-Analyse können sowohl Erkenntnisse für das operative Geschäft als auch für künftige Anforderungen gewonnen werden.

Nicht alle Anwender wollen SQL verstehen, sondern wollen schnell Antworten auf Ihre Fragen, um ihre eigentlichen Kernaufgaben im Unternehmen zu erfüllen, benötigen Informationen. Um Daten lesen und verarbeiten zu können, braucht man technisches Wissen über SQL und Datenmodelle, erst dann entstehen wertvolle Analysen. Rollenverständnis und Data Governance sind 2 wesentliche Aspekte, um erfolgreich Self Service BI im Unternehmen anzuwenden.

Die IT sollte ich darauf konzentrieren, die Schnelligkeit und Qualität ihres Lösungsportfolio´s stetig zu verbessern. Die Fachseite sollte Data Governance als wichtigen Bestandteil von Business Intelligence anerkennen. Gemeinsam kann den Anwendern dadurch ermöglicht werden, ihre Kernaufgaben im Unternehmen wahrzunehmen.

Datawarehouse - Trends bei Datenbanktechnologien

Aufgrund der steigenden Datenmenge und den höheren Ansprüchen an Anzahl und Dauer von Zugriffen, entwickeln sich neue Datenbanktechnologien, die als Alternative zu relationalen DBMS eingesetzt werden können. Wichtig ist hierbei, die Eigenschaften dieser neuen Technologien im Kontext des Gesamtsystems, in dem die Datenbank ein Bestandteil ist, zu verstehen.

<u>In Memory Datenbanken</u>

Abgekürzt IMDB, halten Daten primär im Hauptspeicher eines Rechners um dadurch schnellere Zugriffszeiten auf die gespeicherten Daten zu ermöglichen, da Hauptspeicher i.d.R. wesentlich höhere Zugriffsgeschwindigkeiten und effizientere Zugriffsalgorithmen als Festplatten vorweisen können.
Jedoch gehen bei einem System Ausfall Daten im Hauptspeicher verloren. Ein wesentliches Merkmal von Datenbanken ist Transaktionskonsistenz, die durch Persistenz der Daten erreicht wird. IMDB's stellen dazu folgende Methoden bereit:

[2] A **sandbox** is a testing environment that isolates untested code changes and outright experimentation from the production environment or repository. Quelle: http://en.wikipedia.org/wiki/Sandbox_%28software_development%29

- Zustände der Daten werden in Zeitintervallen erfasst und auf persistente Speichermedien geschrieben (Snapshots). Daten, die zwischen diesen Zeitintervallen anfallen, können verloren gehen.
- Transaktionsprotokolle werden ausgelesen und auf persistente Speichermedien geschrieben (Replikation).
    - Bei einer asynchronen Replikation werden Transaktion und Persistenz getrennt, was eine hohe Performanz und auch ein hohes Risiko des Datenverlust mit sich bringt.
    - Bei einer synchronen Replikation werden Transaktion und Persistenz zusammengeführt, was bei Schreibzugriffen zu längeren Laufzeiten führen kann, dafür hohe Datensicherheit durch eine ACID[3] konforme Transaktionssteuerung sicherstellt.
- Einsatz von NVRAM[4] Speicher, der bei Systemausfällen den letzten Datenzustand wieder herstellen kann
- Hybride Ansätze, welche die oben genannten Methoden kombinieren um damit ACID und Hochverfügbarkeit bei gleichzeitig schnellen Datenzugriffen zu erreichen.

NoSQL Datenbanken

Not only SQL Datenbanken[5] sind strukturierte Datenspeicher die keine relationalen Algorithmen, wie Datenmodelle in der 3. Normalform, verfolgen und teilweise auch auf SQL oder ACID verzichten. NoSQL Datenbanken skalieren horizontal und arbeiten verteilt, damit werden Ziele wie schnelle Zugriffe, Hochverfügbarkeit oder niedrige Hardwarekosten erreicht.

Folgende Ansätze werden als NoSQL Datenbanken bezeichnet:

- *Key-Value* Paare sind einfache Lookup Strukturen, wobei die Schlüssel je nach Datenbank Hersteller auch gruppiert oder erweitert werden. Diese können In Memory oder auf Festplatte gespeichert werden.

[3] siehe auch http://de.wikipedia.org/wiki/ACID
[4] NVRAM oder "Non-Volatile Random-Access Memory", sind Hauptspeichermodule mit eigener Stromversorgung.
[5] CAP und BASE sind weitere, zentrale Eigenschaften des NoSQL Ansatzes. Die Begriffe können in Wikipedia nachgelesen werden.

- *Columnar* speichern Daten spaltenorientiert und besitzen die Eigenschaften von NoSQL. Dies unterscheidet diesen Ansatz von den relational spaltenorientierten Datenbanken[6].
- *Document* verwenden auch den Key-Value Ansatz, wobei der Wert ein Dokument darstellt.
- *Graph* modelliert die Verbindungen zwischen Datenobjekten, welche wiederum als Key-Value Paare dargestellt sind.

NoSQL Datenbanken sind für einfache, schnell zugängliche Datenstrukturen bei gleichzeitig hohen Datenvolumen und einer hohen Anzahl an Zugriffen optimiert. Es gibt wenig Restriktionen bei Datenmodellen, Änderungen sind damit auch bei großen Datenmengen leicht umsetzbar.

Relational spaltenorientierte Datenbanken

Daten werden spaltenorientiert in Blöcke geschrieben, mit dem Ziel weniger IO zu generieren und damit einen schnelleren Zugriff auf Daten zu erreichen. Konventionelle RDBMS schreiben Daten zeilenorientiert in Blöcke.

Beim Auslesen der Daten werden im spaltenorientierten RDBMS nur die Spalten gelesen, die in der Query enthalten sind, bei den zeilenorientierten RDBMS werden Zeilen und damit alle Spalten gelesen. Dies kann bei der hohen Spaltenzahl in Starschemata einen entscheidenden Unterschied darstellen. Zudem komprimieren spaltenorientierte RDBMS besser als zeilenorientierte, da direkt auf komprimierte Daten zugegriffen wird, ohne diese über die CPU zu dekomprimieren. Durch interne Indizes wird ein schneller Zugriff bei Filter mit n>1 Spalten erreicht, oft werden CPU Algorithmen direkt vom RDMBS effizient eingesetzt. Spaltenorientierten RDBMS generieren damit oft weniger I/O als zeilenorientierte RDMBS, insbesondere bei Reporting und Analyse.

Eigenschaften der Relational spaltenorientierten Datenbanken:

- spaltenorientierte Speicherung
- Komprimierung der Daten
- effizienter CPU Einsatz

[6] siehe unten

- Sortierung der Daten
- hohe Zugriffsgeschwindigkeiten

Folgendes Beispiel verdeutlicht die Speicherung, wobei am Bonus auch die Komprimierung nachvollzogen werden kann.

ID	Nachname	Bonus
1	Müller	7000
2	Meier	3000
3	Berg	3000

Abbildung 1: Beispiel

Zeilenorientiert

- 1 Zeile wird zusammenhängend in n>0 Blöcke gespeichert.
- 1,Müller,7000; 2,Meier,3000; 3,Berg,3000

Spaltenorientiert

- 1 Spalte wird zusammenhängend in n>0 Blöcke gespeichert, wobei die Daten sortiert nach Zeilen abgespeichert werden.
- 1,2,3; Müller,Meier,Berg; 7000,3000,3000

Im Unterschied zu NoSQL Columnar enthalten die Relational spaltenorientierten Datenbanken eine wichtige Eigenschaft, sie bilden Daten relational ab. Dies unterstützt insbesondere Auswertungen und Analysen mit konventionellen SQL, auch die meisten Reporting und Analyse Anwendungen unterstützen SQL.

Big Data

Big Data bedeutet große Datenmengen in unterschiedlichsten Datenformaten (Bilder, Dokumente, Geodata usw.) ausreichend schnell zugreifbar und auswertbar zu halten. Man spricht auch von den 3 Vs, "Volume, Variety and Velocity". Relational Spaltenorientierte Datenbanken, NoSQL Datenbanken, Massive Parallele Systeme, flexible Datenschemata[7] sind Bestandteile zur Technologie des Big Data.

Weitere Trends

[7] Beim speichern der Daten muss nicht zwingend ein Datenschema vorliegen. Datenmodelle werden oft erst bei der Auswertung von Daten generiert.

Weitere, hier nur stichwortartig angesprochene Trends sind:

- Predictive Analytics: Computer gestützte, statistische Vorhersagemodelle
- Mobile BI: Verteilung von Analysen und Reports auf mobile Endgeräte
- Business Intelligence 3.0.: Kollaboration, interaktive und leicht bedienbare Reportingsysteme, Vorhersagemodelle, Cloud

Fazit

Business Intelligence wird bei zunehmender Dynamik in der Geschäftswelt und einer stetig steigenden Datenmenge immer wichtiger in Unternehmen. Der Artikel hat anhand der 2 Themen Business Intelligence und Datawarehouse einige Trends vorgestellt, mit der diese Dynamik und Datenmenge gemeistert werden können.

Klaus Rohrmoser beschäftigt sich seit über 10 Jahren mit Datenbank Lösungen, Anforderungsanalyse und BI. Er ist Gründer von data2fact, ein Unternehmen das Business Intelligence und Datawarehouse Lösungen für den Mittelstand anbietet. Erfahren Sie mehr auf www.data2fact.de.

Kontakt:

Klaus Rohrmoser

klaus.rohrmoser@data2fact.de