

Von außen kommend hat man es bei der Beobachtung von Systemen und Anwendungen immer leicht. Dann wundert man sich über diese und jene Eigentümlichkeit. „Warum machen die das so und nicht anders?“ oder „Es wäre doch so einfach, wenn ...“

Einblicke in die Praxis gewachsener Data-Warehouse-Systeme

Alfred Schlaucher, ORACLE Deutschland B.V. & Co. KG

Die Bedingungen, unter denen die Data-Warehouse- und Business-Intelligence-Systeme aufgebaut und verwaltet werden, sind meist von außen nicht erkennbar. Schon deswegen sollte man sich mit Bewertungen zurückhalten. Aber von außen kommend, hat man auch eine gewisse Distanz und läuft nicht Gefahr „den Wald vor lauter Bäumen nicht mehr zu sehen“. Deshalb sind die Data-Warehouse-Reviews von Oracle auch eine so interessante Quelle von Aspekten aus der gelebten Praxis. Der Startpunkt ist ein standardisierter Fragebogen, den man in einem lockeren Gespräch zwischen Systemverantwortlichen und Oracle-Mitarbeitern beantwortet. In der Regel münden die Gespräche dann nach Stunden in intensive Diskussionen über das „Für und Wider“ von Techniken und Verfahren. Nicht selten fällt der Satz: „Das haben wir schon immer so gemacht.“ Spätestens jetzt ist ein Ansatz für Verbesserungen gefunden, denn etwas Verfestigtes ist meist schon überholt.

Ansprüche und Ziele von Data-Warehouse-Verantwortlichen

Unbestritten gehören Analyse-Systeme heute zu den Top-Anwendungen. Entsprechend heftig ist demzufolge auch die Diskussion um Technologien, Architekturen und Tools. Zwei Hauptziele sind in den Diskussionen erkennbar:

- Die Flexibilität der Informationen soll gesteigert werden. Gemeint ist damit die Verzahnung von Daten aus unterschiedlichen Sachgebieten sowie eine variable Detailtiefe, von

grob aggregiert bis auf die Ebene operativer Daten.

- Man will in immer kürzeren Zeitabständen aktuelle Informationen bekommen, also im Extremfall die geschäftliche Entwicklung bis kurz vor Berichterstellung sehen.

Unausgesprochen stehen auch immer die Kosten im Raum. Das Senken von Kosten wird nicht als Hauptziel genannt. Sie schränken jedoch den Handlungsspielraum ein und entscheiden als gesetzte Zwangslage bei der Umsetzung von Ideen meist mit.

Die Ziele sind nicht neu, aber die Wege und Irrungen dorthin. Nicht selten verselbstständigt sich beispielsweise ein spezieller Tool-Einsatz und versperrt den Weg zu einer optimierten Anwendung. Einmal getroffene Entscheidungen zu revidieren, fällt dann schwer.

Bei der Herangehensweise an Data-Warehouse-Aufgaben und auch bei der Bewertung über Sinnhaftes und Unsinniges nutzt Oracle meist ein entsprechendes Schichtenmodell (in der Regel „Inmon“). Die Anzahl der Schichten und die Menge der Schritte sind dabei weniger entscheidend. Wichtig ist die Lösung der beiden Hauptaufgaben:

- Das technische, aber auch fachliche Zusammenführen von Informationen aus mehreren Sachgebieten
- Die Informationen so bereitzustellen, dass sie auch ein Fachfremder versteht

Die Rahmenbedingungen sind so zu wählen, dass Datenflüsse möglichst kurz und redundanzfrei bei geringstem Hardware-Einsatz sind. Hinter diesen

pauschalen Formulierungen steckt ein komplettes Konzept, das der Autor auf Anfrage gerne diskutiert.

Immer wieder „Stove Pipes“

Vielen Systemen fehlt der sogenannte „Enterprise-Bezug“ beziehungsweise ist die „Enterprise-Fähigkeit“ oft nicht ausgeprägt genug. Damit meint man die Möglichkeit, über eine einzige Abfrage Informationen aus unterschiedlichen und über das Aufgabenspektrum des Unternehmens verteilte Sachgebiete zu erfahren. Offenbar sind die Systeme zunächst nur für ein Sachgebiet entwickelt worden (einen Data Mart). Daten wurden direkt aus den Vorsystemen in die Auswerte-Modelle geladen, ohne sie zuvor zu granularisieren und vergleichbar zu machen. Als die Anforderungen nach weiteren Informationen hinzukamen, wurde erweitert, indem man einfach nur einen weiteren Data Mart neben den bestehenden stellte, ohne die vorhandenen Informationsobjekte (und auch ETL-Prozesse) miteinzubeziehen. Es entstand ein System ohne einheitlichen Informationshaushalt, aber mit vielen autarken Data Marts.

Diesen auch „Stove Pipes“ genannten Effekt findet man gerade in großen Warehouse-Systemen in größeren Unternehmen. Die Folgen werden leider erst Jahre nach der Erstimplementierung sichtbar. Nachteilig ist hier:

- Die fehlende Flexibilität für die Endanwender, die keine zusammenhängenden und stimmigen Informationen erhalten.
- Die hohen Kosten für die Weiterentwicklung, wenn das System er-

weitert werden muss. Oft muss man komplett neue ETL-Strecken und Data Marts entwickeln.

Eine teure Neuimplementierung des gesamten Systems ist oft die einzige Lösung. Es ist erstaunlich, wie wenig das Drei-Schichten-Architektur-Prinzip in der Praxis diskutiert und konsequent angewendet wird.

Starre Implementierung des Schichtenmodells

Es gibt aber auch das entgegengesetzte Extrem: Eine sachgebietsübergreifende, zentrale Warehouse-Schicht ist zwar vorhanden, sie ist aber gegenüber den Auswertemodellen (Data Marts) schon fast hermetisch abgeschottet: „In mein zentrales Data Warehouse kommt kein Anwender rein.“ Wenn jedoch kein Anwender hineinkommt, dann muss man eben alle benötigten Daten in die Data Marts kopieren, auch wenn die Tabellen extrem groß sind und auch wenn die Daten bei dem Weg in die Data Marts nicht mehr modifiziert werden, also keine Mehrwerte erhalten.

Solche Systeme erzeugen unnötigen ETL-, Hardware- und Verwaltungs-Aufwand. Agilität (schnelles Handeln bei Neuanforderungen) wird erschwert.

Aus Gründen der Ressourcen-Schonung kann man folgende Regel formulieren: „Keine „1:1“-Kopien im Data Warehouse“. Das bedeutet, dass große Bewegungsdaten- und Fakten-Tabellen nur einmal in dem System vorzuhalten sind und Referenzdaten, die sich (im Gegensatz zu Stammdaten) nicht beziehungsweise kaum verändern, ebenfalls nur einmal vorhanden sein sollten.

Dies gelingt nur, wenn eine zentrale Warehouse-Schicht auch für Endbenutzer-Zugriffe geöffnet bleibt. Das bedeutet auch, dass Dimensionen über Schichtengrenzen hinweg eine Fakten-Tabelle referenzieren, die aus Kostengründen nur einmalig in der zentralen Warehouse-Schicht liegt und nicht im Data Mart.

Auch wenn es manchem DWH-DBA widerstrebt: Alle Schichten sollten nur logisch als Schicht begriffen werden und physisch nicht getrennt (etwa auf unterschiedlichen Rechnern / Datenbanken) liegen.

Entfremdung von Data Warehouse und Business Intelligence

„Fachanwendungen (BI) für die Fachanwender in der Fachabteilung und Data Warehouse in die IT-Abteilung.“ Je größer ein Unternehmen, desto mehr verfestigt sich diese Vorgehensweise.

Es entsteht folgende Situation: Neben dem zentralen und von der technischen IT verwalteten Data Warehouse gibt es mehrere und zum Teil voneinander unabhängige Business-Intelligence-Anwendungen mit unterschiedlichen Teams und oft auch Tools. Die Zentral-IT beschränkt sich auf die Service-Leistung „die Bereitstellung von Data-Warehouse-Daten /-Schnittstellen“ auf Anforderung. Das erfolgt über einen formalen Prozess. Bei ungenügenden Ergebnissen (falsche Daten, nicht zusammenhängend, nicht die nötige Detailtiefe etc.) erfolgen neue, zeitaufwändige Anforderungen an die IT, bis die Anwender resignieren und eigene Wege gehen.

Diese Vorgehensweise führt zu vielen Sonderwegen, Sonderzonen und Sonderverantwortlichkeiten, letztlich zu doppelter Arbeit, mehrfachen Ressourcen und leider oft auch nicht abgestimmten Kennzahlen. Während sich die IT in der technischen Optimierung der Datenbank austobt, verheddern sich die Fachanwender in einem Wust nicht abgestimmter Einzeldaten, die sie oft nicht mehr verstehen. Und allzu oft glaubt man, durch die Anschaffung eines neuen BI-Tools mehr Flexibilität zu schaffen, etwa durch Zugriffe auf Daten der Vorsysteme und schnelle In-Memory-Bearbeitung. Teilweise werden auch ETL-Aufgaben in die BI-Tools verlagert. Die Anforderung nach Flexibilität für die Endanwender ist die Legitimation. Der Datenaufbereitungsprozess wird zum Glücksspiel.

Richtig wäre es, den gesamten Datenfluss von den Vorsystemen bis hinein in die Berichte/Analysemodelle als einen zusammenhängenden Informationsbeschaffungs-Prozess zu verstehen und diesen Prozess durch eine Hand verwalten/modellieren zu lassen. Dies setzt voraus, dass in der IT mehr über die fachlichen Anforderungen (Informationsbedarfe) bekannt ist und in den Fachabteilungen mehr

Kenntnisse über die IT-Systeme beziehungsweise existierenden Datentöpfe (Quellen/Vorsystem) vorhanden sind. Helfen würde sicher schon eine bessere Zusammenarbeit zwischen den Data-Warehouse-Verantwortlichen und den Fachabteilungen.

In kaum einem beobachteten Data Warehouse konnten die Verantwortlichen spontan erklären, welche Daten für welche Zielgruppe im Data Warehouse gespeichert sind. Es existieren kaum Daten-Inventare und kaum Listen über die Benutzergruppen mit ihren Daten-Interessen. Eine Liste von zugelassenen Datenbank-Benutzern reicht nicht. Nur wenn man die Interessen und Nutzungsgewohnheiten der Anwender kennt und permanent überprüft, kann man die DWH-Informationen passgenau bereitstellen.

Fehlende Planung des ETL-Prozesses

Sind die Schichten und Datenmodelle im Data Warehouse schon zufällig, so sind es erst recht die ETL-Prozesse. Vor allem, wenn ETL-Tools zum Einsatz kommen, verselbstständigt sich dieser Bereich. Dann ist zu sehen, dass ETL-Entwicklung, Datenmodellierung und Informationsbedarfs-Planung auf unterschiedlichen Schreibtischen stattfinden und nicht Hand in Hand gehen.

Aufgrund fehlender Ausrichtung an einem Schichtenmodell und Informationsbedarfs-Planung entsteht oft ein chaotischer Wirrwarr von einzelnen Lade-Jobs und ETL-Strecken. In den meisten Data-Warehouse-Systemen wird zu viel und zu umständlich geladen. ETL-Prozesse könnten mit weniger Redundanzen, mit geringeren Laufzeiten, mit weniger Hardware und weniger Volumen laufen, wenn man sie besser planen würde und diese Aufgabe nicht nur den ETL-Tool-Spezialisten überlassen würde.

Nachgelagerter Stellenwert des Data Warehouse

Man sollte die besondere Bedeutung eines Data Warehouse für heutige Unternehmen nicht mehr betonen müssen. Aber es gibt noch zu viele Verantwortliche, die das nicht wissen: „Ohne DWH geht vieles in den Unternehmen einfach nicht mehr.“ Dieses Unwissen

herrscht oft in Rechenzentren und in den administrativen IT-Bereichen vor, wo es um Hardware, Storage und Administration geht, Datenbank-Administration nicht ausgeschlossen.

Data-Warehouse-Systeme stehen in der Wahrnehmung nicht an der operativen Front des Unternehmens, es sind Sekundär-Systeme. Sie werden nachgelagert hinter vielen OLTP-Anwendungen irgendwie auch noch mitbetreut. Die technische Administration des DWH wird über die normale DBA-Gruppe erledigt, die allzu oft keinen Unterschied zwischen einer OLTP- und einer DWH-Datenbank macht. Aber: Ein Data Warehouse ist kein OLTP-System. Diese einfache Aussage muss auch in den Rechenzentren und der IT-/DB-Administration berücksichtigt werden. Das bedeutet dedizierte Ressourcen und Verfahren für ein Data Warehouse:

- Separater Storage
- Separate Rechner
- Separate Netze
- Separates Backup-Verfahren

- Separate DBA-Administration (zumindest sollten DBAs über die Besonderheiten in einem DWH Bescheid wissen)

Schwacher Einsatz von Datenbank-Techniken

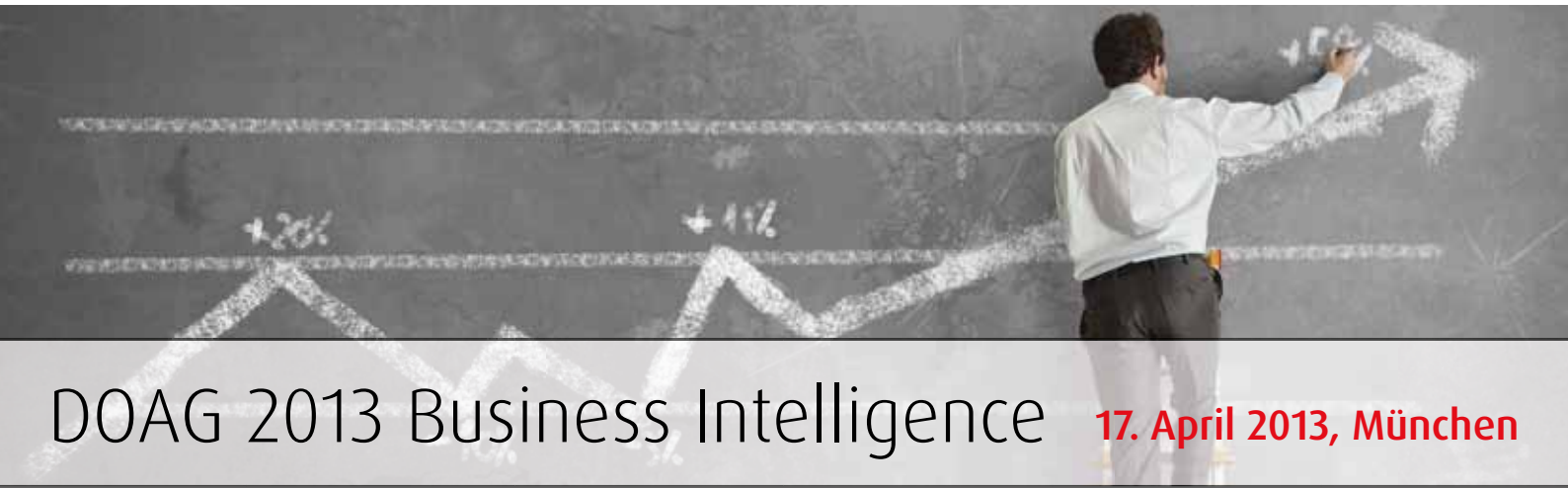
Wichtige DWH-Features der Datenbank werden immer noch zu wenig eingesetzt. Nach einer Schätzung kommen in der Praxis zum Einsatz:

- 70 Prozent Partitioning
- 40 Prozent Materialized Views
- 30 Prozent DWH-spezifische Indizierung (Bitmap, Star Transformation)
- 30 Prozent analytische Funktionen
- 10 Prozent Mining
- 10 Prozent OLAP
- 50 Prozent ETL-Funktionen

Gerade die Features „Materialized Views“, „DWH-Indizierung“ und „analytische Funktionen“ könnten mehr eingesetzt werden. Der Grund für diese Zurückhaltung liegt zum Teil in der Organisation der Teams. Da es sich

um technische Features der Datenbank handelt, fallen sie meist in das Aufgabengebiet eines OLTP-orientierten Datenbank-Administrators. Der Nutzen eines Materialized-View-basierten Kennzahlensystems oder analytisch aufbereiteter Werte ist jedoch nur für Fachanwender erkennbar. Diese wissen wiederum oft nicht um die technischen Möglichkeiten der Datenbank und akzeptieren eine auch nur spartanisch bereitgestellte Lösung. Beide Gruppen müssten stärker zusammenarbeiten.

Alfred Schlaucher
alfred.schlaucher@oracle.com



DOAG 2013 Business Intelligence 17. April 2013, München

Eine Konferenz rund um die Themen Business Intelligence und Data Warehousing

- Themen:
- Data Management, Datenqualität, Data Warehouse, Big Data
 - BI Technology, OBIEE, Oracle BI Suite
 - Advanced Analytics, Data Mining, Exalytics
 - Methodology & Modelling, Agile DWH Systeme

Aussteller:



Quest Software
is now a part of Dell



Kooperationen:



<http://bi.doag.org>

