Den Nutzwert der Datenbasis mit Datamining erhöhen

Detlef E. Schröder
Oracle Deutschland B.V. & Co KG
Hamburg

Schlüsselworte
Datamining, deskriptive Statistik, SQL, Decision Tree, SVM, k-mean, Korrelation, DWH, Datamart, ETL

Einleitung
Das Datawarehouse hat sich in den vergangenen Jahren etabliert. Es kann aber mit wenigen, einfachen Mitteln in seinem Nutzen weiter gesteigert werden. Dies wird anhand der in der Datenbank vorhandenen Möglichkeiten gezeigt und beschreibt somit einen effizienten und effektiven Umgang mit der Oracle Datenbank.

Klassische Ziele und Architektur zur Sicherung des Nutzens eines DWH

Seit Jahren hat sich eine Architektur im DWH etabliert, die die Hauptziele und den Nutzen eines DWH sicherstellt. Dazu gehören vor allem :

- die Zentralität eines Warehouses, d.h. die Loslösung der Daten aus ihrem Entstehungsort und Speicherung in einem neuen Zusammenhang.
- die Historisierung, d.h. über die Lebensdauer in den operativen Vorsystemen hinausgehende Speicherung der Informationen an dieser zentralen Stelle.
- die Umwandlung und qualitative Sicherung der Informationen und der damit verbundenen Generierung von Mehrwerten in einem DWH, im Gegensatz zu den einzelnen Vorsystemen und
- die Loslösung der Informationen aus ihrem Herkunftszusammenhang und damit der Neutralität der Informationen, so das sie dann in andere Zusammenhänge integriert werden und neue Sichten und Zusammenhänge entstehen können.

Diese Regeln an den Entwurf eines DWH haben dazu geführt, das die verschiedenen Informationsbedürfnisse der Anwender, die herkunftsübergreifende oder herkunftsunabhängige Informationen benötigen, befriedigt werden können. Die Informationsbedürfnisse werden in der Regel in der Datamartschicht zusammengestellt und aufbereitet. Hier werden in der Regel auch Aggregationen und Kennzahlen berechnet und in den Datamarts erzeugt.

Diese Berechnungen bedienen sich aber in der Regel der „normalen“ Mathematik und gehen über die Grundrechenarten selten hinaus. Dies hat natürlich vor allem seinen Grund darin, dass die Anwenderbedürfnisse dies auch nicht erfordern. Aber auch, da man weitergehende Statistik nicht im DWH macht, sondern dazu teure Tools einkauft oder sich des Exports nach Excel bedient, um dort seine eigenen Statistik zu betreiben. Diese Entwicklung ist auch dadurch geschehen, dass die Sprache zwischen den DWH Entwicklern und den statistischen Anwendern zu unterschiedlich ist und das Wissen nicht verbreitet wurde. Dieses Hoheitsdenken hat sich in den letzten Jahren noch verstärkt.

Grundsätzlich gibt es aber keinen besseren Ort als Statistik und Datamining zu betreiben als in dem DWH, in der Datenbank, direkt an und mit den Daten. Diese stehen hier in qualitativ gesichertem Zustand zur Verfügung und alle Informationen sind abgreifbar und so stehen dann auch die statistischen und Datamining Ergebnisse, wie Clustering und Segmentierung, zentral zur Verfügung und können für die weiteren Analysen und Auswertungen verwendet werden. Dazu braucht es keinen externen „Rechenknecht" der dies erledigt. Damit werden auch nicht zwei Schnittstellen erzeugt (Export der für die Kalkulation notwendigen Daten und Import der Berechnungsergebnisse). Dies impliziert ja auch Fragen nach Sicherheit und Integrität der Daten, die dort mit hineinspielen, auf die aber in diesem Zusammenhang nicht näher eingegangen wird.

Deswegen wollen wir uns im Folgenden die Möglichkeiten ansehen, die von normalem SQL ausgehend, über das Oracle Datamining bis hin zu einem Ausblick auf Oracle R Enterprise, direkt an und mit den Daten möglich sind und wie dies hilft den Nutzen und den Wert Ihres DWH zu steigern.

Statistische Informationen mit „einfachem" SQL

Über die letzten Versionen sind immer mehr statistische Funktionen und Möglichkeiten direkt ins SQL gewandert. Die Analytischen Funktionen, z.B. Lag() für Periodenvergleiche, Avg() für Gleitende Durchschnitte, ratio_to_report als Anteile und einige weitere, die im SQL Reference Guide und dem Data Warehousing Guide beschrieben sind. Diese beziehen sich aber fast ausschließlich auf die Abfrage-Seite und werden für Berichts-Kennzahlen verwendet.

Es sind allerdings im SQL auch noch weitere Funktionen enthalten, die uns in der statistischen Betrachtung der Inhalte unterstützen. Zum Beispiel wenn wir uns mit der Verteilung von Werten innerhalb einer Spalte beschäftigen und hier mehr Informationen für die weitere Verwendung erhalten wollen. Über die üblichen Werte wie Mittelwert, avg(betrachtete Spalte) hinaus gibt es auch die einfache Anwendung des gewichteten Durchschnitts und auch eine einfache Behandlung von NULL Werten, da ein Count der Spalte die NULL Werte nicht mit rechnet ein count(*) aber schon.

```
SELECT avg(nummerische Spalte, im folgenden NS),
               sum(NS) / count (*),
               sum(NS) / count (NS)
FROM tabelle
```

In dem DBMS_STAT_FUNCS Paket sind weitere Funktionen der deskriptiven Statistik zusammengefasst, die einen ausreichenden Überblick über den Inhalt einer Tabelle zusammenstellen können. Dazu zählen dann auch die Quantile, die Standardabweichung und Six Sigma Werte. Damit läßt sich einfach ein Bericht erzeugen, der über die betrachtete Tabelle eine umfassende Auskunft gibt und damit Aufschlüsse, welche weiteren „verborgenen" Zusammenhänge zu finden sind. Weitere Informationen finden Sie im Oracle Database PL/SQL Packages and Types Reference Guide.

Dazu lassen sich dann, noch immer mit „normalem" SQL, die Korrelationen zwischen zwei Spalten ermitteln. Wie ist also der Aussagewert der Spalten zu betrachten, hängen sie inhaltlich zusammen oder sind sie, von ihrer Aussagekraft her, getrennt zu sehen. Gerade wenn man im folgenden Kluster bilden möchte oder Segmente bilden, helfen diese Informationen die „richtigen" Spalten mit in die Untersuchung einzubeziehen und nicht alles verwenden zu müssen. Die Funktion CORR gibt es auch in verschiedenen Varianten, die den Einsatz flexibel möglich macht.

```
select CUST_ID, CUST_MARITAL_STATUS, CUST_YEAR_OF_BIRTH, CUST_CITY_ID,
  corr(CUST_YEAR_OF_BIRTH, CUST_CITY_ID) over (PARTITION BY
  CUST_MARITAL_STATUS) as correlation
from SH.CUSTOMERS
where CUST_YEAR_OF_BIRTH is not null
  and CUST_CITY_ID is not null
order by CUST_MARITAL_STATUS, CUST_ID;
```

Natürlich gibt es noch einige weitere Funktionen, die auch über die beschreibende Statistik hinaus gehen, wie Regressionsfunktionen und verschiedenen Statistische Testverfahren (Chiquadrat Test, F-Test, t-Test, ...).

Gerade im DWH, wo die Daten bereinigt und geprüft vorliegen, machen diese Untersuchungen und vorbereitenden Analysen Sinn. Da dies auch von dem Rechenaufwand her nicht unerheblich sein kann. Je nach der Inputmenge und der verwendeten Funktionen, ist es erst recht sinnvoll, dies direkt mit den Daten an den Daten zu machen und diese nicht erst zu bewegen. Das Warehouse kann hier ein Lieferant weiterer Informationen sein, die den Analysten und Datamining Entwickler wichtige Hilfsmittel und Kenngrößen zusätzlich zur Verfügung stellt. Dies steigert den Wert nicht nur der Informationen die im DWH zusammen getragen sind sondern nutzt auch die vorhandene Infrastruktur für das Warehouse um diese Informationen zu ermitteln.

Datamining

Die bisher gesammelten Informationen und Ergebnisse der deskriptiven Statistik helfen bei den nun folgenden Aufgaben des Datamining. Hier kann uns der SQL Developer, in den der Oracle Data Miner integriert ist, weiter helfen. Hier sind die wesentlichen Aufgaben des Datamining zusammengefasst und stehen unter einer Oberfläche zur Verfügung.

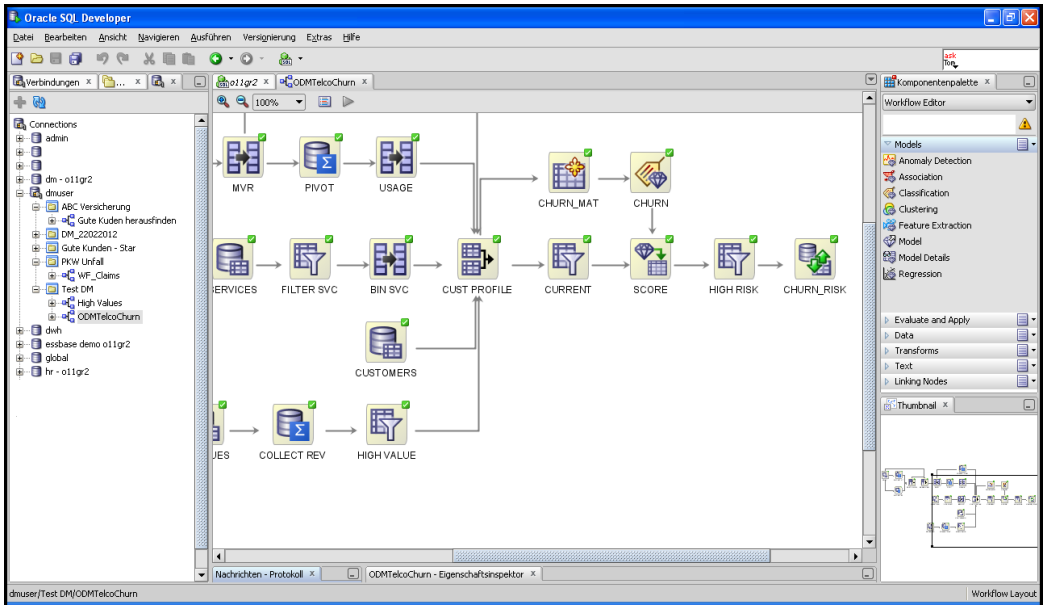

Abb. 1: Beispiel für den Oracle Dataminer im SQL Developer

Zu diesen Aufgaben gehört die weitere Aufbereitung der Daten, um sie mit den Datamining Methoden und Mittel weiter verarbeiten zu können. Natürlich kann gerade diese Aufgabe, wenn sie denn regelmäßig durchgeführt werden muss, mit in den ETL – Prozess (Extraktion, Trasformation und Laden) des DWH integriert werden und sollten auch dort stattfinden. Gerade wenn es um die weitere Bereinigung der Daten geht, wie z.B. Umgang mit NULL Feldern, fehlenden oder unvollständigen Informationen. Dies stellt eine Kernaufgabe des DWH da und sollte dann auch in die dort bestehenden Prozesse integriert werden und nicht, wie sonst auch üblich, zu neuen Werkzeugen mit neuen Schnittstellen, geleitet werden. Auch wenn sich eine Datamining Struktur, eine sehr breite Tabelle in der die zu untersuchenden Dimensionen und die Fakten inklusive Aggregationen aufgelöst sind, von der, in der Datamartschicht normalerweise verwendeten Star Schemata, grundsätzlich unterscheidet, sollten diese Strukturen doch gerade aus der oben beschrieben Kern DWH Schicht einfach abzuleiten sein und dann mit in der Zugriffschicht / Datamartschicht des DWH angeboten werden. Diese werden im Oracle Dataminer dann auch als Bestandteile aufgerufen und in einem Workflow zusammengestellt.

Es werden wieder keine Daten bewegt, sondern nur vorbereitet und erst zur Laufzeit des Workflows abgearbeitet. Die Ergebnisse können dann zurückgeschriebene Werte sein, z.B. der Kluster, in dem sich eine Element befindet, oder aber eine Funktion, die ich immer wieder auf diese oder neu hinzukommende Werte anwenden kann. Letzteres führt dann zu einer operationalisierten Verwendung des Dataminings.

Da für die Erreichung einer Aufgabe, z.B. des Clusterings von Kunden, verschiedene Verfahren angewendet werden können, bietet der Oracle Data Miner die Hilfe an, alle in Frage kommenden Verfahren, anzuwenden und dann Anhand von statistischen Kenngrößen, das Verfahren auszuwählen, dass für diese Datengrundlage und diese Aufgabe am Besten geeignet ist. Trotzdem hat der erfahrene Statistiker aber immer noch die Möglichkeit gezielt einzugreifen und auch einzelne Verfahren mit gezielter Parametrisierung zur Ausführung zu bringen.

Für die Aufgaben stehen folgende Algorithmen zur Verfügung:
- Decision Tree (DT)
- Naive Bayes (NB)
- Generalized Linear Models (GLM)
- Support Vector Machine (SVM)
- Minimum Description Length (MDL)
- Apriori (AP)
- k-Means (KM)
- Non-Negative Matrix Factorization (NMF)
- One Class Support Vector Machine (One-Class SVM)
- Orthogonal Partitionining Clustering

Auf OTN gibt es für das Datamining einfache und übersichtliche Beispiele, die einen guten Einstieg in den Oracle Data Miner und das Datamining ermöglichen.

(http://www.oracle.com/technetwork/database/options/advanced-analytics/odm/index.html)

Und auch hier kann man mit dem Datamining wieder den Informationsgehalt und die Verwendbarkeit der Daten, in der Datenbank, in dem DWH weiter steigern. Es braucht kein neuer Datenhaushalt geschaffen werden und dank spezieller Transformations – Views, können verschiedene Schritte und

Aufgaben ohne Materialisierung durchgeführt werden. So entsteht eine schnelle und übersichtliche Lösung.

Ausblick auf Oracel R Enterprise

In den letzten Jahren hat sich im Datamining Umfeld ein Werkzeug etabliert, das sich gerade mit der Visualisierung einen Namen gemacht hat - R.

R ist ein Desktop Tool, das bei vielen Datamining Entwicklern durch die breiten Möglichkeiten der Statistik und der umfangreichen und fast endlosen grafischen Aufbereitung Einzug gehalten hat.

Der Nachteil dieses Werkzeuges ist allerdings die Einschränkung auf einen Prozess mit dem Client-Hauptspeicher. Hier hat Oracle mit dem Oracle R Enterprise angesetzt. So ist eine Verwendung von R aus der Datenbank heraus möglich. R Skripte werden in PL/SQL Rutinen eingebaut und können so von der Datenbank verwendet und verwaltet werden. Dies geschieht dann unter der Kenntnis und Verwendung des Wissens um die Struktur und Möglichkeiten der Daten in der Datenbank, z.B. dem Partitioning. Darüber hinaus kann die Datenbank dann mehrere R Prozesse auf dem Datenbankrechner, der meist mit genügend Hauptspeicher versehen ist, aufbauen, um die parallelen Prozesse in der Datenbank mit Ergebnissen zu versehen.

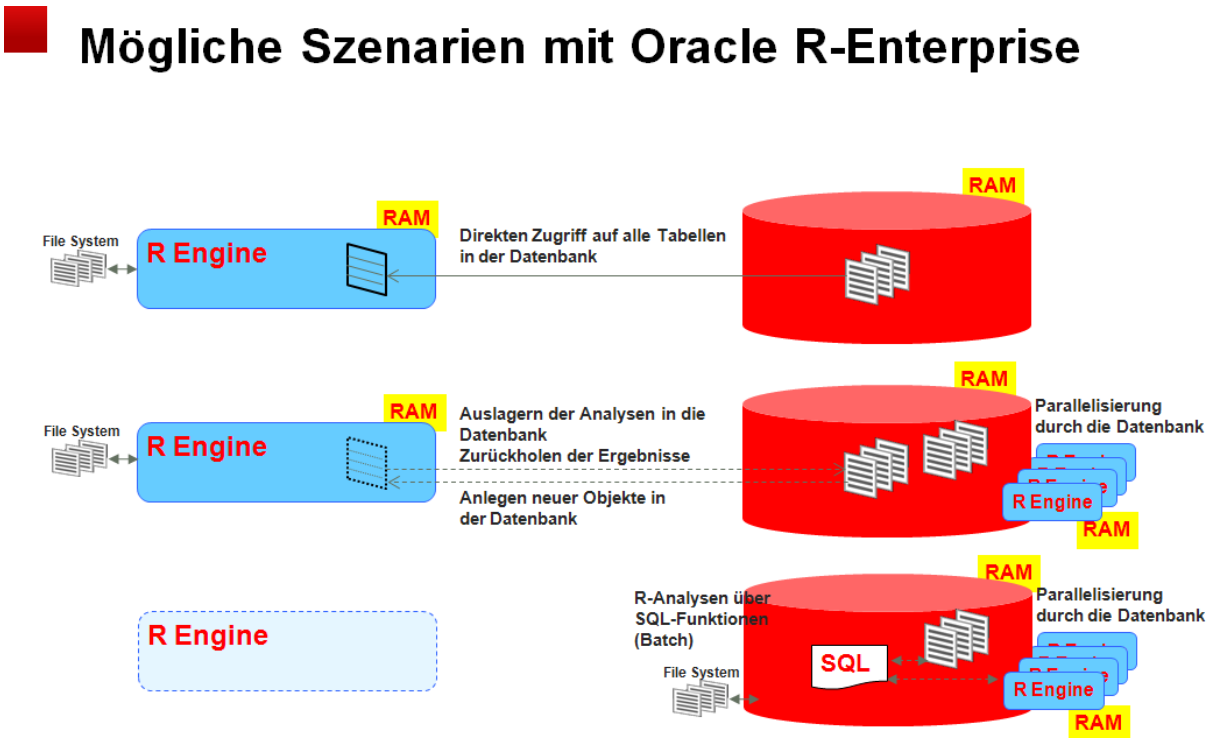

Abbildung 2: Mögliche Einsatzszenarien für Oracle R Enterprise

Dies soll aber nur einen Ausblick darstellen, wie die Informationen aus dem DWH auch mit diesen Möglichkeiten, die auch Bestandteil der Oracle Advanced Analytic Option sind, zu der auch das Oracle Datamining gehört, bereichert werden können und dies den Nutzen des DWH steigen kann.

Zusammenfassung und Schluss

Da im DWH schon die Informationen so bereinigt und passgenau vorliegen, bietet es sich geradezu an, hier die Grundlage für einen weiteren Nutzen über die Anwendung von Statistik zu schaffen und das ohne den Verlust von Sicherheit und Performance aus der bestehenden Umgebung heraus. Dies führt zu einem weiteren Anwendungsgebiet für die Daten des DWH und zu einer breiteren Nutzung dieser Informationen.

Kontaktadresse:
Detlef E. Schröder
Oracle Deutschland B.V. & Co KG
Kühnehöfe 5
D-22761 Hamburg

Telefon: +49 (40) 8 90 91 - 423
E-Mail Detlef.E.Schroeder@oracle.com
Internet: www.oracle.de