Guided Navigation basierend auf SAP Netweaver BIA

Bernhard Jäksch, Robert Lembke, Barbara Stortz,
Stefan Haas, Anja Gerstmair, Franz Färber

SAP Netweaver BI - Walldorf
{b.jaeksch, r.lembke, barbara.stortz, stefan.haas, anja.gerstmair, franz.faerber}@sap.com

Kurzfassung: Interaktives Online Analytical Processing mit klassischen Operationen wie Drill-Down, Roll-Up und Pivoting gehören mittlerweile zum Standard-Repertoire von Analysewerkzeugen. Der Benutzer wird dabei auf sogenannten Drillpfaden (meistens als Hierarchien und Dimensionen bezeichnet) durch den Datenbestand geführt. Das Prinzip des „Guided Navigation" eröffnet eine alternative Art durch Datenbestände zu navigieren und Zusammenhänge zwischen einzelnen Dimensionen innerhalb eines Datenwürfels zu erkennen. Der Nachteil des Guided Navigation aus Sicht der Datenbanksysteme liegt jedoch in einer im Vergleich zum klassischen OLAP signifikant höheren Aggregationsleistung in allen Dimensionen pro Benutzerinteraktion. In dieser Demo wird ein System vorgestellt, welches das Prinzip der Guided Navigation auf Benutzeroberfläche umsetzt und basierend auf dem SAP Netweaver BIA die Aggregationsleistung erfüllt, die notwendig ist um ein interaktives Analyseverhalten auf großen Datenbeständen zu ermöglichen.

1 Einleitung

Die Definition von Hierarchien auf dimensionalen Strukturen stellt aktuell den Status Quo aller OLAP-Werkzeuge zur interaktiven Analyse von großen Datenbeständen dar. Entlang vordefinierter Drillpfade wird es dem Benutzer erlaubt, Datenbestände zu verfeinern (Drill-down) oder höhere Aggregate zu berechnen (Roll-Up). Der zentrale Nachteil dieser Interaktionsmethode besteht darin, dass der Benutzer vorab erahnen muss, hinter welchem Aggregat sich ein interessanter Zusammenhang versteckt um eine Drill-Down-Aktion auf dieses Aggregat auszulösen.

Das Prinzip der Guided Navigation geht einen alternativen Weg und ermöglicht dem Benutzer eine aufgefächerte Sichtweise auf die Dimensionen eines Datenwürfels. Die Navigation erfolgt dabei nicht mehr nur ausschließlich über vordefinierte Hierarchien, sondern über mögliche Zusammenhänge in den Fakt-Daten. Der Nutzwert der Guided Navigation wird insbesondere durch eine Aussage von Joseph Busch (http://www.taxonomystrategies.com) mit folgender Aussage unterstrichen: "Four independent categories [facets] of 10 nodes each can have the same discriminatory power as one hierarchy of 10,000 nodes." Konkret soll folgendes Beispiel, welches ebenfalls im Rahmen der Demo zur Veranschaulichung des Analyseprinzips und der dahinter stehenden Datenbanktechnik benutzt wird, das allgemeine Prinzip nochmals verdeutlichen:

Das Beispielszenario reflektiert einen Online-Shop für Bekleidung (e-Fashion) mit dimensionalen Attributen wie Jahr, Monat und Produktinformationen auf Artikel-, Gruppen- und Kategorienebene. Als Kennzahlen sind in dem multidimensionalen Schema Umsatzzahlen, etc. definiert, die auch gegen eine Geschäftsdimension (City, State-Ebene) ausgewertet werden können. Das Schema, d.h. die Kennzahlen und die zur Guided Navigation zur Verfügung stehenden Attribute werden in sogenannten InfoSpaces definiert. Abbildung 1 gibt einen Einblick in die Auswahl und Definition von InfoSpaces für das konkrete e-Fashion-Szenario.

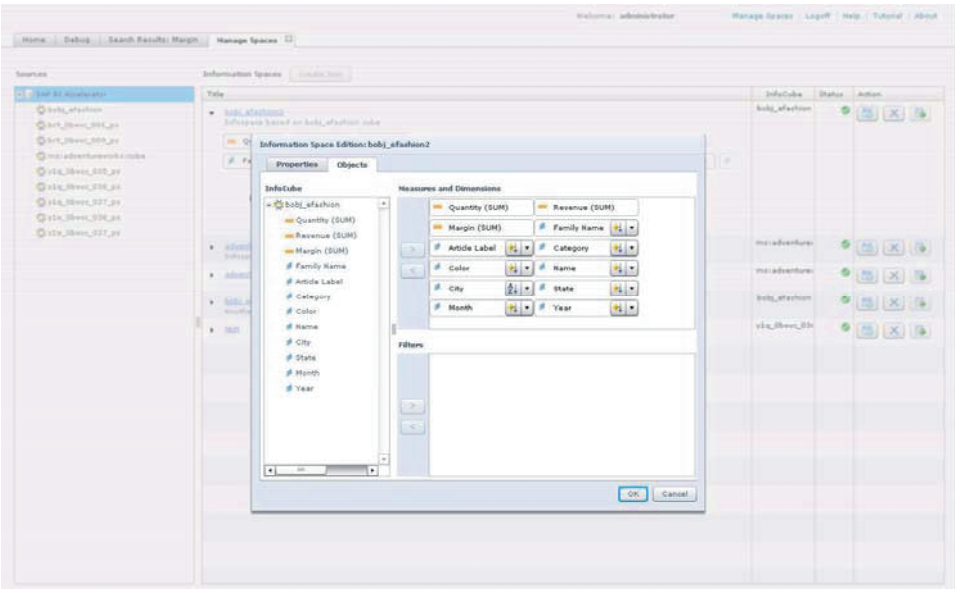

Abbildung 1: Auswahl und Definition von InfoSpaces

Das Prinzip der Navigation ist in Abbildung 2 dargestellt. Dabei werden Einschränkungen auf einzelnen beschreibenden Attributen vorgenommen (Jahr 2006, 2007). Das System ermittelt im Hintergrund zwei Ergebnisaspekte. Zum einen werden die dadurch ausgewählten Kennzahlen durch eine (normale, OLAP-übliche) Aggregation ermittelt und graphisch in unterschiedlichsten Formen dargestellt. Zum andern – und dies ist der wesentliche Vorteil einer Guided Navigation - werden nur die Ausprägungen in den verbleibenden Dimensionen ermittelt, die tatsächlich über mindestens einen Faktdatensatz zur den ausgewählten Dimensionsbereichen aufweisen. Im konkreten Szenario bleiben nur die Werte in den nicht von einem Benutzer selektierten Attributen (wie beispielsweise Produktfamilie) erhalten, die mindestens einen Verkauf in den Jahren 2006 und 2007 aufweisen konnten. Die Guided Navigation unterstützt somit nicht nur eine explizite Selektion durch den Benutzer (wie im OLAP üblich) sondern auch eine implizite Selektion durch eine Auswertung der Beziehung über die Faktdaten. Insbesondere bei dünnbesetzten Datenwürfeln hilft diese Art der Navigation enorm.

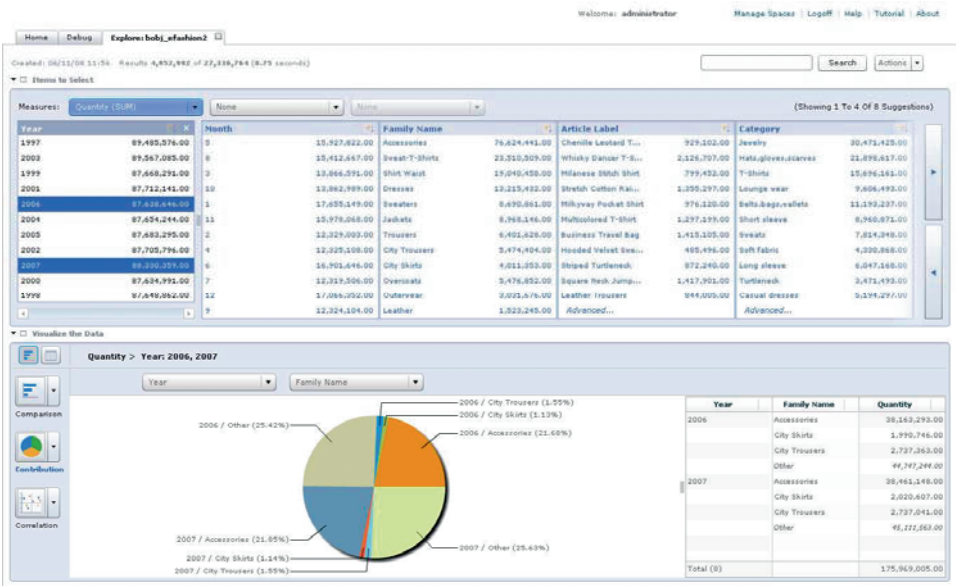

Abbildung 2: Auswahl und implizite Selektion über Faktdatensätze

Die Herausforderung für die Datenbankseite besteht nun darin, bei jeder Benutzerinteraktion pro beteiligtes Attribut zu bestimmen, wie oft eine Beziehung zum eingeschränkten Datenwürfel noch existiert. Jede Interaktion impliziert dadurch eine Gruppierung pro Dimensionsattribut; bei hoch-dimensionalen Würfel müssen entsprechend viele Aggregationsanfragen abgesetzt werden.

3 Datenbanktechnologie

Das Prinzip der Guided Navigation wird basierend auf der Datenbank-Engine SAP Netweaver BIA demonstriert. Dabei handelt es sich um ein hoch-paralleles System, welches relationale Star- und Snowflake-Schemata spaltenorientiert organisiert. Um eine hohe Anfrage-Performance zu erreichen werden die Daten auf der einen Seite zusätzlich zur vertikalen Spaltenaufteilung noch horizontal partitioniert und die dadurch entstandenen Partitionen auf unterschiedliche Rechnerknoten allokiert um eine maximale Parallelisierung von Datenbankanfragen zu erzielen. Das Produkt wird bis zu einer Größe von standardmäßig 32 Rechnerknoten mit je 2 CPU/4cores (Harpertown) vorkonfiguriert ausgeliefert. In einer Laborumgebung in Zusammenarbeit mit dem Hardware-Partner IBM wurde eine lineare Skalierung bis zu 140 Knoten nachgewiesen (http://www.sap.com/platform/netweaver/pdf/BWP_BI_Accelerator_WinterCorp.pdf).

Auf der anderen Seite verfolgt der SAP Netweaver BIA die Philosophie der Hauptspeicherdatenbanken, in dem der gesamte Datenbestand in komprimierter Form im Hauptspeicher gehalten werden kann. Zugriff auf Disk erfolgt nur beim initialen Befüllen der Hauptspeicherstrukturen nach dem Start. Die unterschiedlichen Komprimierungsverfahren basieren auf einem verzeichnisbasierten Ansatz, wobei die

eigentlichen Werte in einem Dictionary gehalten werden. Die Daten selbst, d.h. die Einträge einer Spalte einer Relation, bestehen dann entsprechend nur noch aus einer Sequenz von Codes, die auf den jeweiligen Eintrag im Verzeichnis verweisen. Der SAP Netweaver BIA selbst läuft des Weiteren als spezifische Konfiguration des SAP TREX Projektes, in welchem neben der BIA-Engine weitere Module spezifische Aufgaben wie Textsuche, Text-Mining etc. übernehmen können. Abbildung 3 gibt einen Überblick über die Komponenten des Gesamtsystems.

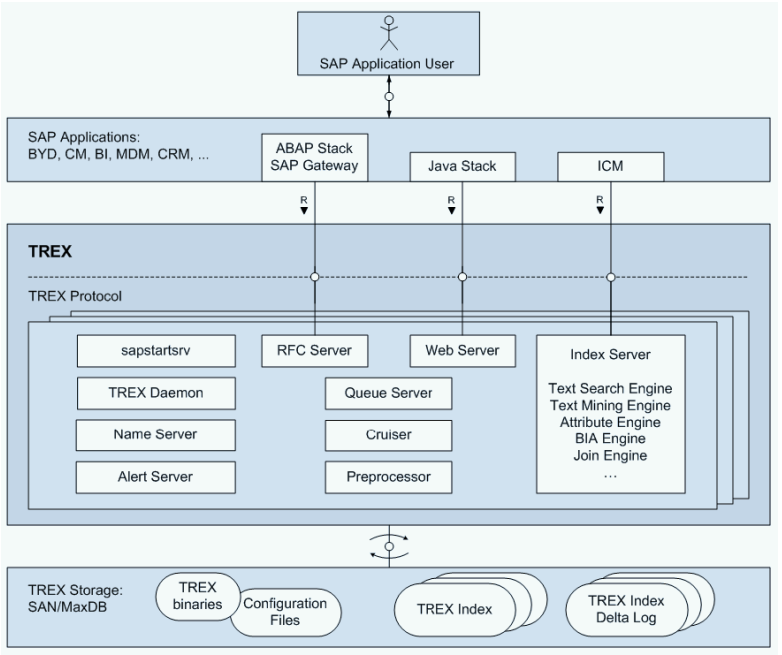

Abbildung 3: Architektur der TREX-Infrastruktur

2 Demonstration

In der Demo wird zum einen das User-Interface für eine Guided Navigation gezeigt. Dabei handelt es ich um eine vollständig Browser-basierte Benutzeroberfläche mit dem Komfort einer Rich Client-Applikation Zum anderen wird die Technologie des SAP Netweaver BIA erklärt. Der BIA ist eine spaltenorientierte Daten-Analyse-Engine, die verteilt über mehrere Rechnerknoten Datenbestände in komprimierter Form im Hauptspeicher hält und parallele Aggregationsoperation ausführt. Im Rahmen der Demo wird ein BIA auf einem lokalen System gezeigt. Zusätzlich besteht (bei Internetverbindung) die Möglichkeit auf große Datenbestände (>1Mrd Fakten) zuzugreifen und die Skalierbarkeit des BIAs durch die hochgradige Interaktion zu demonstrieren. In der Diskussion mit den Interessierten Besuchern werden entsprechende Konzepte (Komprimierung, Parallelisierung, etc.) erläutert, die diese Form der Interaktion ermöglichen.