Op zoek naar 'dimensionele intelligentie'

Hoeveel sterren krijgt uw rapportagetool?

Frank Habers en Peter Dieleman

Het gebruik van een dimensioneel model voor een datawarehouse biedt vele voordelen. De prestatie is het bekendste en meest genoemde wapenfeit. Hoewel dit inderdaad niet te onderschatten valt, is er een ander, minstens even belangrijk winstpunt. Dimensioneel modelleren biedt een framework van standaards, methoden en technieken voor het ontwerpen van datawarehouses, met alle voordelen van dien. Rapportagetools zouden goed binnen dit raamwerk moeten passen. Frank Habers en Peter Dieleman beschrijven van welke hoge klasse deze tools moeten zijn.

De methode van dimensioneel modelleren geeft onder meer standaardoplossingen voor het omgaan met de historie van attributen (*slowly changing dimensions*), het laden van het datawarehouse, het modelleren van heterogene dimensies en de omgang met aggregatietabellen. Steeds meer -maar nog lang niet alle- leveranciers gebruiken dit framework als uitgangspunt bij het introduceren van nieuwe functionaliteit.

Een goed voorbeeld daarvan is de *starjoin*. Voor het behalen van prestatievoordelen wordt bij een starjoin eerst een cartesisch product gemaakt van de geselecteerde dimensietabellen, alvorens de centrale feitentabel (die vele malen groter is) te benaderen. Dit kan enorme prestatievoordelen opleveren, omdat pas in de laatste stap de grootste tabel wordt benaderd. De meeste databaseleveranciers hebben zich inmiddels gebaseerd op deze standaard.

Volwassen datawarehousing, tevreden gebruikers

In 1995 werd het boek *The Data Warehouse Toolkit* van Ralph Kimball uitgegeven, waarin hij het fenomeen *dimensioneel modelleren* introduceerde. Sindsdien is het razendsnel gegaan met dimensioneel modelleren en inmiddels is Kimballs methode algemeen geaccepteerd als dé manier van modelleren voor datawarehouses. Maar een dimensioneel model werkt niet zonder meer goed. Om alle voordelen ervan optimaal tot hun recht te laten komen, moeten de rapportagetools die op het datawarehouse worden ingezet, voldoen aan bijzondere eisen. Deze gaan verder dan de functionaliteit die we inmiddels mogen verwachten van de diverse rapportagetools, zoals het gebruik van een semantische laag tussen database en gebruiker, een gebruikersvriendelijke grafische interface en drill-functionaliteit. Met deze eisen wordt de lat voor de diverse leveranciers weer een paar centimeter hoger gelegd, en groeit de datawarehousemarkt verder naar volwassenheid. Uiteindelijk gaat het natuurlijk om het resultaat: kunnen gebruikers (eenvoudig en snel) de vragen stellen die zij willen stellen?

Rapportagetools[1] zouden op een zelfde manier voordelen kunnen behalen doordat de makers deze baseren op het framework. De twee belangrijkste zaken nemen we daarbij onder de loep:

- intelligentie - hoe 'slim' is het tool ten opzichte van het dimensionele model;
- presentatie - hoe goed kan het tool het dimensionele model presenteren?

EEN VOORBEELD-MODEL

De gewenste functionaliteit beschrijven we aan de hand van het voorbeeld in figuur 1. Het is een eenvoudig dimensioneel model uit een bancaire omgeving, waarin naast de transacties (zoals periodieke af- en bijschrijvingen en transacties via geld- en betaalautomaten) van klanten het rekeningsaldo maandelijks wordt vastgelegd (ofwel een stand gegeven). De fictieve gebruiker Kim Ball wil op dit model analyses uitvoeren.

INTELLIGENTIE

Dimensies zijn eigenlijk niets meer dan de *constraints* (beperkingen) bij het stellen van een vraag. Het bepalen van deze beperkingen doet Kim Ball veelal stapsgewijs, zoals veel gebruikers doen.

'Dimensional browsing'

Kim wil bijvoorbeeld eerst een selectie maken uit een lijst van regio's. Zij kiest regio West en wil nu een keuze kunnen maken uit de klanten in regio West, waar-

bij zij -vanzelfsprekend- niet de klanten uit andere regio's wil zien. Deze manier van exploreren wordt ook wel *dimensional browsing* genoemd.

▶ *Een rapportagetool moet in staat zijn op basis van een eerste selectie een tweede selectie te beperken.*

Let wel: deze functionaliteit wordt gebruikt voordat de uiteindelijke vraag van de gebruiker op het dimensionele model wordt uitgevoerd. Het biedt de gebruiker de mogelijkheid snel zicht te krijgen op de inhoud van een dimensie, waardoor hij snel de juiste vraag te formuleren.

'Multipass SQL'

Kim wil nu van de geselecteerde klanten het aantal transacties en het saldo rapporteren. Dit betekent dat beide feitentabellen (transacties en saldi) benaderd moeten worden (zie figuur 1). Eén van de uitgangspunten bij een dimensioneel model is dat nooit meer dan één feitentabel wordt gecombineerd in één SQL-commando, maar daarmee wil Kim vanzelfsprekend niet worden lastig gevallen.

▶ *Het rapportagetool moet deze situatie allereerst herkennen -detecteren dat twee feitentabellen benaderd moeten worden-, vervolgens twee SQL-commando's genereren, de resultaatsets combineren en ten slotte tonen in één rapport.*

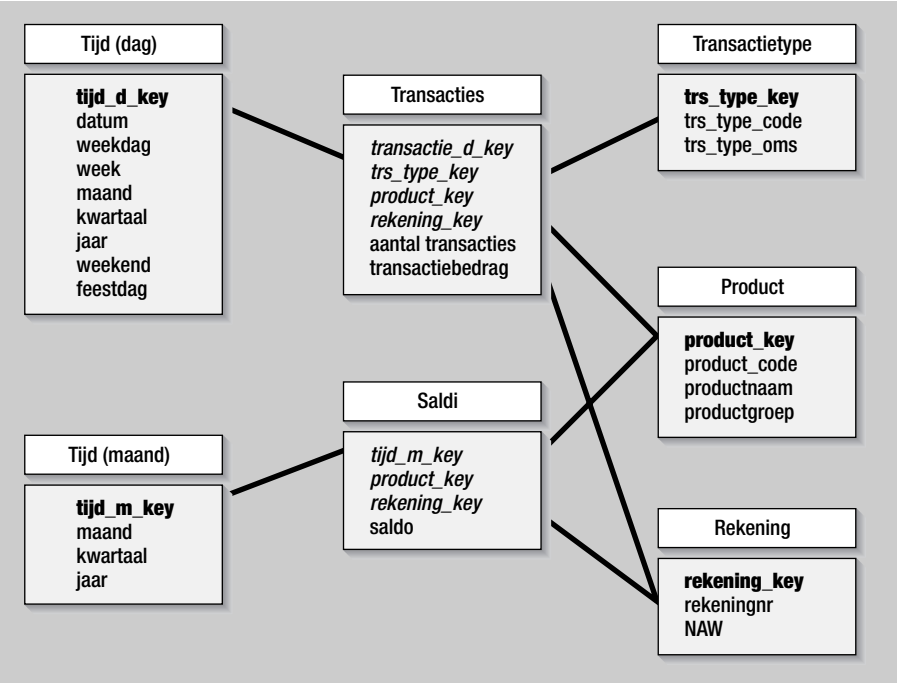
Dit principe wordt *multipass SQL* genoemd. Als het rapportagetool de optie heeft voor benadering van een dimensioneel model, is deze functionaliteit relatief eenvoudig te ontwikkelen.

Herkenning sterschema's

Vervolgens wil Kim per klant het saldo en het aantal transacties per transactietype zien. In figuur 1 zien we dat de dimensie transactietype alleen geldig is voor de transactiefeiten en niet voor de saldigegevens; het saldo kan niet naar transactietype worden opgesplitst.

▶ *Het rapportagetool zou een optie moeten hebben om Kim te waarschuwen dat deze dimensie niet geldig is voor saldigegevens.*

Het tool moet in elk geval voorkomen dat een combinatie wordt gemaakt van de tabellen transactietype, klant, transactiefeiten en saldifeiten, ofschoon deze in de semantische laag (indirect) gekoppeld zijn. Dit zou immers foutieve resultaten opleveren. De semantische laag moet één sterschema daarom herkennen als een groep van tabellen die bij elkaar horen. De vraag kan dan in twee rapporten worden beantwoord, namelijk het saldo per klant en het aantal transacties per klant, per transactietype.



Figuur 1: Eenvoudig dimensioneel model uit een bancaire omgeving.

Applicatieniveau

Kim is nog steeds niet uitgevraagd en wil nu het verschil zien tussen het aantal transacties dat per acceptgiro en dat welk via de betaalautomaat is uitgevoerd. SQL is niet sterk in het maken van dit soort vergelijkingen. Om deze vraag te beantwoorden moet het rapportagetool op applicatieniveau functionaliteit bevatten voor het uitvoeren van vergelijkingen. Dit is een eenvoudig voorbeeld, maar een businessanalist als Kim wil natuurlijk ook kijken naar optellingen, percentages, ratio's, vergelijkingen in de tijd, cumulatieven, afwijking ten opzichte van een gemiddelde enzovoort.

Nog lang niet alle leveranciers gebruiken het framework

▶ *Een rapportagetool zal een krachtige rekenmodule op applicatieniveau moeten bevatten, die de resultaten van de relatief eenvoudige SQL-commando's kan combineren en bewerken.*

'Outerjoin'

Het antwoord op haar vorige vraag heeft Kims interesse gewekt voor klanten die de afgelopen maand géén transacties hebben uitgevoerd. Zij maakt via de tijddimensie een selectie op de huidige maand en geeft aan dat alleen die klanten getoond moeten worden bij wie geen transacties horen. Deze vraag kan eenvoudig worden opgelost door in de semantische laag van het rapportagetool de relatie tussen de tabel Klant en de tabel Transactiefeiten als een *outerjoin* te definiëren. Dit zou echter betekenen dat altijd een outerjoin wordt uitgevoerd, wat in de meeste gevallen tot performanceproblemen leidt.

▶ *De oplossing hiervoor is de mogelijkheid per vraag (of per dimensie) in te stellen of een outerjoin moet worden uitgevoerd. Kim moet de mogelijkheid krijgen om dit op eenvoudige wijze te doen.*

Vanzelfsprekend kan zij daarbij beter uit de voeten met de vraag: "Wilt u ook de klanten zien zonder transacties?", dan met: "Moet een left outerjoin worden uitgevoerd?". De vraag is gebaseerd op de

businessnaam van de specifieke tabel. Deze optie moet per dimensietabel kunnen worden ingesteld.

Semi-additieve meetwaarden

Kim gooit het nu over een andere boeg en wil het gemiddelde totaalsaldo van regio West zien van de afgelopen zes maanden. In deze regio worden 10.000 bankrekeningen beheerd, en de totale optelling van saldi wordt gedeeld door 60.000 (6 x 10.000). Dit levert echter niet het gewenste resultaat op, want het totaal moet gedeeld worden door 6 (maanden). Dit komt doordat de meetwaarde saldo *semi-additief* is. Saldi van verschillende maanden kun je niet optellen, maar saldi van bijvoorbeeld verschillende klanten wel.

▶ *Het rapportagetool moet weten dat deze meetwaarde niet optelbaar is via de tijddimensie en altijd gemiddeld moet worden over deze dimensie. Per meetwaa*rde moet je kunnen vastleggen via welke dimensies deze kan worden geaggregeerd.

Tussenresultaten opslaan

Met al haar opgedane kennis vordert Kim in haar analyses. Zij wil nu graag de klanten zien die een bovengemiddeld saldo hebben, die bovendien een toename in het aantal transacties hebben in de afgelopen twee jaar van meer dan vijftig procent, die meer dan twee bankproducten hebben aangeschaft en die in de afgelopen vijf jaar ten minste eenmaal zijn verhuisd. Het mag duidelijk zijn dat deze vraag niet valt op te lossen via één SQL-commando.

▶ *De vraagstelling moet worden opgesplitst in een aantal eenvoudige query's. De tussenresultaten hiervan worden bij voorkeur opgeslagen in de database, zodat het resultaat van de ene query, de basis (of het filter) kan zijn voor een andere.*

Dit betekent wel dat het rapportagetool nu een aantal extra tabellen in de database nodig heeft om een vraag te kunnen beantwoorden, en dat levert uiteraard meer beheerinspanning op.

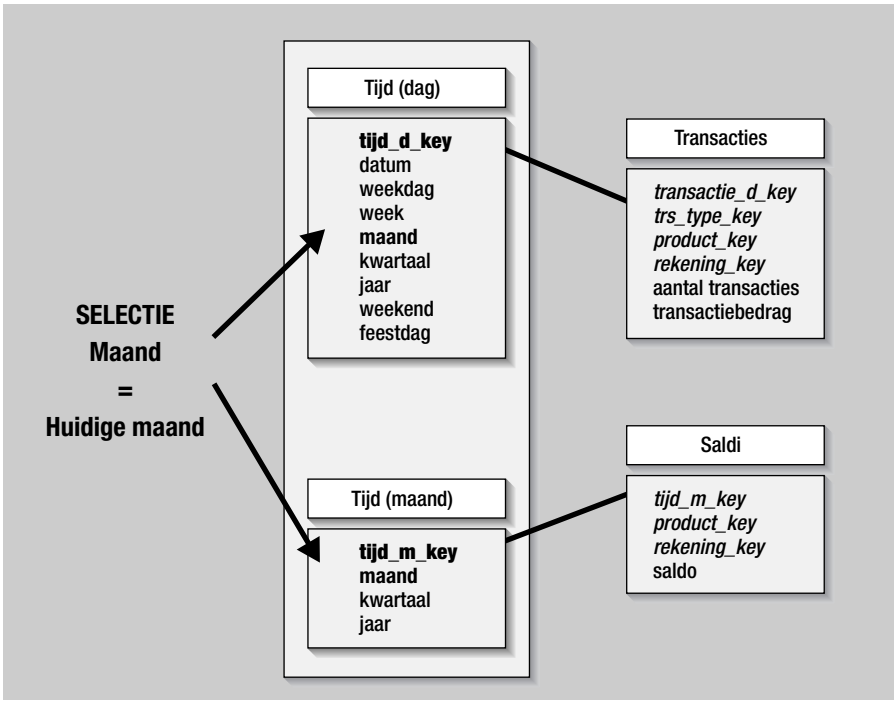
Er wordt nooit meer dan één feitentabel gecombineerd in een SQL-commando, maar daarmee wil de gebruiker niet lastig worden gevallen

Kims vraag is een typisch voorbeeld van de vragen die voorkomen in (marketing)-omgevingen, waarin men veel analyses doet ten behoeve van cross-selling, up-selling en retentie. Dit soort vragen is vooral gericht op het maken van allerlei combinaties van groepen van producten en klanten op basis van geselecteerde kenmerken. Tussenresultaten opslaan is hierbij essentieel.

Geconformeerde dimensies

Van schrik stelt Kim een wat simpelere vraag: zij wil van de afgelopen maand het gemiddelde saldo zien en het gemiddelde transactiebedrag rapporteren. Daarvoor dienen beide feitentabellen te worden benaderd. De selectie vindt plaats vanuit de tijddimensie op dagniveau (voor de transacties) en de tijddimensie op maandniveau (voor het saldo). Dit zijn *geconformeerde dimensies*, waarbij de laatstgenoemde dimensie gegevens op geaggregeerd niveau (maand) bevat. Dit wordt gevisualiseerd in figuur 2. Kim raakt nu in verwarring, omdat zij in haar vraagstelling tweemaal de afgelopen maand moet selecteren.

▶ *Het zou natuurlijk vriendelijker zijn voor Kim als het gebruikte hulpmiddel één filter voldoende vindt, en vervolgens de gebruiker vraagt of dit filter van toepassing is op beide meetwaarden.*



FIGUUR 2: SELECTIE VANUIT GECONFORMEERDE DIMENSIES.

PRESENTATIE

Tot zover de beschrijving van de 'intelligentie' die rapportagetools moeten bevatten voor een goed gebruik van een dimensioneel model. Een ander belangrijk aspect van een dergelijk hulpmiddel is de visualisatie en presentatie van het dimensionele model. Dit aspect is nogal onderbelicht in de literatuur over dimensioneel modelleren. Harm van de Lek heeft dit onderwerp in zijn artikelen wel onder de loep genomen door een nieuw principe te introduceren: OASI (One Attribute Set Interface). Dit principe maakt ons onder meer duidelijk dat het fysieke dimensionele model niet dezelfde structuur hoeft te hebben als het aan de gebruiker gepresenteerde model.

Een duidelijk voorbeeld hiervan is de minidimensie Demografie uit het voorbeeld, waarin op basis van de postcode van een klant allerlei demografische kenmerken zijn opgenomen: welstand, inkomen enzovoort.

Ook aandacht voor cardinaliteit

Een nadeel van het dimensionele model is dat de semantiek binnen een dimensie verdwijnt: met name de cardinaliteit en de kenmerken van attributen. De cardinaliteit tussen attributen (bijvoorbeeld: Woonplaats kan een of meer klanten bevatten) binnen een dimensie, en ook tussen dimensies, moet eenvoudig toonbaar zijn aan een gebruiker. Dit houdt onder meer in dat de hiërarchie van een dimensie inzichtelijk gemaakt moet worden aan de gebruiker. Vanzelfsprekend kunnen meerdere hiërarchieën per dimensie bestaan. Daarnaast moet het onderscheid tussen analyse-attributen (zoals woonplaats, product en jaar) en beschrijvende attributen (zoals straatnaam, telefoonnummer en transactienummer) duidelijk zijn voor de gebruiker. De beschrijvende objecten zal Kim immers niet gebruiken om te analyseren (de omzet per telefoonnummer is in het voorbeeld niet erg interessant), maar wel om de uitkomst van een analyse te verrijken ("geef mij het telefoonnummer van klanten die voldoen aan mijn selectiecriteria"). Rapportagetools zouden dit onderscheid moeten kunnen aanbrengen met verschillende iconen en verschijningsvormen van de attributen.

▶ *Hoewel Demografie in het fysieke model een aparte dimensie vormt, is het voor de gebruiker begrijpelijker als de demografische kenmerken in het gebruikersmodel binnen de klantdimensie worden getoond.*

Zo kunnen we meer presentatie-eisen stellen aan rapportagetools die op een dimensioneel model worden ingezet.

METADATA

Eén van de onderdelen van het standaard framework dat het dimensionele model geeft, is de implementatie van *slowly changing dimensions*, op verschillende manieren. Datamodelleringstools hebben inmiddels functionaliteit die dit concept ondersteunt. Voor de gebruiker is het natuurlijk belangrijk dat rapportagetools ook duidelijk aangeven of en hoe de historie van dimensiekenmerken wordt vastgelegd. Als Kim een mailing wil sturen naar alle klanten die vorige maand zijn verhuisd, moet zij goed weten of deze historie is opgeslagen en hoe zij deze vraag moet stellen. Daarmee is meteen het onderwerp metadata aangeroerd.

FIGUUR 3: STERSCHEMA MET MEETWAARDEN MET INBEGRIP VAN DE DIMENSIES.

▶ *Een rapportagetool moet bij voorkeur allerlei algemene kenmerken van een attribuut kunnen vastleggen en tonen (definitie, omschrijving, vervuilingsgraad, bronverwijzing enzovoort) en men moet ook eenvoudig kunnen zoeken in deze metadata.*

De gebruiker moet bijvoorbeeld per meetwaarde intuïtief leren welke dimensies geldig zijn op welk niveau. Zoals we in figuur 1 zien, is de dimensie Transactietype niet geldig voor de saldi. In de meeste rapportagetools is dit op te lossen door per feit -of liever gezegd per sterschema- een folder te creëren waarin de meetwaarden zijn opgenomen met inbegrip van de dimensies (zie figuur 3). Op die manier ziet de gebruiker snel welke dimensies hij kan gebruiken voor analyses op meetwaarden.

Helaas zit bij deze oplossing een addertje onder het gras. Zij gaat immers voorbij aan het feit dat onderhoud wordt gepleegd op de attributen. De omschrijving van een attribuut kan wijzigen, of de (afgeleide) definitie kan wijzigen, of de opmaak van het attribuut. Dit onderhoud dient uiteraard bij voorkeur voor ieder attribuut maar eenmaal plaats te vinden. Is een dimensie echter binnen meerdere folders opgenomen, dan moet het onderhoud eveneens op meerdere plaatsen worden uitgevoerd. En dat willen we nu juist voorkomen.

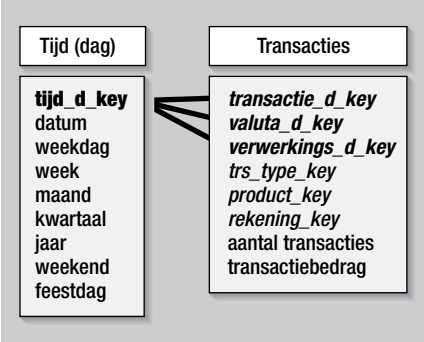
Een simpele, maar effectieve oplossing voor dit probleem is het de mogelijkheid van toevoeging van *verwijzingen* naar het originele object in de semantische laag van het hulpmiddel (vergelijk met de *verwijzing* op het Windows-platform). De verwijzing neemt de eigenschappen over van het originele attribuut, zodat alleen het originele attribuut onderhouden moet worden.

Leveranciers hebben inmiddels voldoende tijd gehad. Toch bezit geen enkel tool alle functionaliteit

Surrogaatsleutels

Ten slotte zien we in figuur 4 dat de dimensie Datum op verschillende manieren wordt gebruikt, namelijk als valutadatum, transactiedatum en verwerkingsdatum. In de figuur zien we dat hiervoor verschillende *surrogaatsleutels* zijn opgenomen. Deze verwijzen fysiek naar dezelfde datumtabel.

▶ *Indien Kim Ball vraagt welke transacties op 1 oktober zijn uitgevoerd (transactie-*



FIGUUR 4: SURROGAATSLEUTELS BIJ UITEENLOPEND GEBRUIK VAN DE DIMENSIE DATUM.

datum) en op 2 oktober zijn verwerkt (verwerkingsdatum), moet het rapportagetool hierop eenvoudig antwoord kunnen geven.

Gelukkig kunnen de meeste tools dit, door gebruik te maken van een *alias*.

TOOLS

Dit artikel is zeker niet volledig in het benoemen van 'dimensionele' functionaliteit die een serieus rapportagetool moet bezitten, maar beschrijft zeker typische situaties die voorkomen bij het gebruik van een dimensioneel model. Toetsen we de momenteel beschikbare tools op de genoemde functionaliteit, dan blijkt het volgende.

Allereerst: een aantal hulpmiddelen biedt heel aardige oplossingen voor een aantal van de genoemde problemen, maar geen enkel tool herbergt alle functionaliteiten in zich. Op zichzelf is dat opmerkelijk, omdat dimensioneel modelleren een breed geaccepteerde methode is. Bovendien heeft Kimball een aantal van de beschreven problemen reeds in 1995 aangestipt. Leveranciers van rapportagetools hebben inmiddels voldoende tijd gehad om de gewenste functionaliteit op te nemen in hun product. Is dit niet het geval, dan is dit een goede indicatie van het innovatievermogen en de visie van de betreffende leverancier. Dat kan een belangrijk argument zijn bij het maken van de keuze voor een hulpmiddel dat bedrijfsbreed voor een lange termijn wordt ingezet binnen een organisatie.

> *Of een rapportagetool voldoet wordt in hoge mate bepaald door de vragen van Kim Ball*

CONCLUSIE

Het is van belang in een vroeg stadium te bepalen wat de 'dimensionele intelligentie' en de presentatiemogelijkheden van een hulpmiddel zouden moeten zijn en in hoeverre deze de voordelen van het dimensionele model versterken. Of een rapportagetool voldoet voor uw organisatie, wordt daarbij in hoge mate bepaald door de eisen en wensen van de gebruikers. Kortom: de vragen van uw Kim Ball. ●

Noten en literatuur

1. Onder rapportagetool verstaan wij in dit kader business intelligence-tools die rapportages samenstellen op basis van SQL.

R. Kimball, *The data warehouse toolkit*, 1995.
R. Kimball, *The data warehouse lifecycle toolkit*, 1998.
H. van der Lek, F. Habers, M. Schmitz, *Sterren en dimensies,* DB/M Essay, 1999.
J. Groff, P. Weinberg, *The complete Reference SQL.*

Drs. Frank Habers (fhabers@inergy.nl) en ir. Peter Dieleman (pdieleman@inergy.nl) zijn als datawarehouse-consultants van Inergy Consult Nederland BV betrokken bij diverse datawarehouseprojecten van grote organisaties.